International Review of the Red Cross (2020), **102** (913), 235–259.
Digital technologies and war
doi:10.1017/S1816383120000466

IRRC

Stepping back from the brink: Why multilateral regulation of autonomy in weapons systems is difficult, yet imperative and feasible

Frank Sauer
Frank Sauer is a Senior Research Fellow at Bundeswehr University in Munich. He serves on the International Panel on the Regulation of Autonomous Weapons and is a member of the International Committee for Robot Arms Control, which co-founded the international Campaign to Stop Killer Robots.
Email: frank.sauer@unibw.de.

Abstract
This article explains why regulating autonomy in weapons systems, entailing the codification of a legally binding obligation to retain meaningful human control over the use of force, is such a challenging task within the framework of the United Nations Convention on Certain Conventional Weapons. It is difficult because it requires new diplomatic language, and because the military value of weapon autonomy is hard to forego in the current arms control winter. The article argues that regulation is nevertheless imperative, because the strategic as well as ethical risks outweigh the military benefits of unshackled weapon autonomy. To this end, it offers some thoughts on how the implementation of regulation can be expedited.

Keywords: artificial intelligence, lethal autonomous weapons systems, arms control winter, regulation, Convention on Certain Conventional Weapons, strategic stability, human dignity.

Introduction

The United Nations (UN) Convention on Certain Conventional Weapons (CCW) is the epicentre of the global debate on autonomy in weapons systems. The CCW's purpose "is to ban or restrict the use of specific types of weapons that are considered to cause unnecessary or unjustifiable suffering to combatants or to affect civilians indiscriminately".[1] In CCW parlance, the weapon autonomy issue is called "emerging technologies in the area of lethal autonomous weapons systems" (LAWS). In November 2019, CCW States Parties decided to, once again, continue their deliberations on LAWS. For the first time, however, these talks, which had previously been conducted between 2014 and 2016 in informal meetings and since 2017 within the framework of an expert subsidiary body called a Group of Governmental Experts (GGE), were mandated to produce a specific outcome. For ten days in 2020 and for an as-yet unknown number of days in 2021 (when the CCW's next Review Conference is due), the GGE was and is tasked with debating and fleshing out "aspects of the normative and operational framework" on LAWS.[2] In addition, in Annex III of their 2019 report, States Parties adopted eleven guiding principles to take into account going forward.[3] After the first five-day meeting of 2020 was postponed and then conducted in a hybrid format due to the current global COVID-19 pandemic, the second meeting had to be shelved, and it is currently unclear when and how the talks can resume.

While some States – most prominently Russia – have displayed no interest in producing new international law in the CCW, arguing that "concerns regarding LAWS can be addressed through faithful implementation of the existing international legal norms",[4] others – such as Germany – claim that nothing short of "an important milestone" has already been reached with the 2019 report cited above, even describing the adopted eleven guiding principles as a "politically binding regulation".[5]

Meanwhile, the international Campaign to Stop Killer Robots (Killer Robots Campaign, KRC) is criticizing CCW diplomacy as "moving forward at a snail's pace", with low ambitions and negligible outcomes despite widespread

1 United Nations Office in Geneva, "The Convention on Certain Conventional Weapons", available at: https://tinyurl.com/y4orq8q5 (all internet references were accessed in December 2020).

2 UN, *Meeting of the High Contracting Parties to 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: Revised Draft Final Report*, UN Doc. CCW/MSP/2019/CRP.2/Rev.1, Geneva, 15 November 2019 (CCW Meeting Final Report), p. 5, available at: https://tinyurl.com/y3gjy7mk.

3 *Ibid.*, p. 10.

4 Russian Federation, *Potential Opportunities and Limitations of Military Uses of Lethal Autonomous Weapons Systems: Working Paper Submitted by the Russian Federation*, UN Doc. CCW/GGE.1/2019/WP.1, 15 March 2019, p. 5, available at: https://tinyurl.com/yx9op3n4.

5 German Federal Foreign Office, "Foreign Minister Maas on Agreement of Guiding Principles relating to the Use of Fully Autonomous Weapons Systems", press release, 15 November 2019, available at: www.auswaertiges-amt.de/en/newsroom/news/maas-autonomous-weapons-systems/2277194.

https://doi.org/10.1017/S1816383120000466 Published online by Cambridge University Press

public opposition to LAWS and some thirty countries (twenty-six of which are CCW States Parties) calling for the immediate negotiation of a new, binding legal instrument rather than continuing talks on frameworks and principles, which the KRC tends to consider vague and redundant respectively.[6]

It should be noted up front that the term LAWS itself is problematic. After all, neither "lethality" nor "autonomy" are decisive factors in the debate. The military application of non-lethal force raises concerns as well (take the prohibition against blinding lasers as just one example), and the term "autonomy", philosophically speaking, inappropriately anthropomorphizes machines that have limited agency and are incapable of reasoning and reflecting, as well as being unable to take on responsibility. Nonetheless, at this point the term LAWS is widely used as a shorthand, so the article will stick to this common vocabulary. Also, the article uses the term "regulation" – rather than, for instance, "ban" – because what potential new, binding international law on this issue is commonly understood to eventually codify is not a prohibition of a category of weapons. Instead, it is a positive obligation to retain meaningful human control over the use of military force. And while one might argue that ensuring meaningful human control and prohibiting autonomous weapons (AKA "killer robots") are two sides of the same coin, these two sides nevertheless represent different ways of approaching the issue, as I will argue further below. Lastly, I use the term "technology diffusion" rather than "proliferation" because the latter suggests a distribution from one or only a few points of departure (as in the case of nuclear proliferation) whereas the former suggests an omnidirectional spread from multiple sources, a more fitting picture in this case of widely (and oftentimes even commercially) available hardware and software.

In what follows, I first explain why it is so challenging for everyone involved in the debate to get a conceptual handle on the issue and, for CCW States Parties, to agree on impactful multilateral regulation on LAWS. I argue that finding proper language and a suitable legal framing for the retention of meaningful human control, in light of the enormous military value ascribed to unshackled weapon autonomy, is what makes regulating LAWS so exceptionally difficult. Subsequently, I discuss the implications of inaction, making the case for why retaining human control over the use of force is indeed imperative due to the strategic and ethical risks outweighing the potential military benefits. Lastly, I put forward some suggestions on how regulation could be advanced in practice, the formidable challenge of gathering enough political will amongst CCW States Parties notwithstanding. This is followed by a brief conclusion.

6 KRC, "Alarm Bells Ring on Killer Robots", 15 November 2019, available *at:* www.stopkillerrobots.org/2019/11/alarmbells/; Richard Moyes, "Critical Commentary on the 'Guiding Principles'", Article 36, November 2019, available at: www.article36.org/wp-content/uploads/2019/11/Commentary-on-the-guiding-principles.pdf.

https://doi.org/10.1017/S1816383120000466 Published online by Cambridge University Press

Why regulating weapon autonomy is difficult: Conceptual pitfalls and power politics

From UN Secretary-General António Guterres to prominent members of the artificial intelligence (AI) and tech communities[7] to most States Parties of the CCW, there is near unanimity that LAWS raise various legal, strategic and ethical questions and concerns.[8] Even so, within the CCW States Parties, a consensus on new, binding international law is still a long way off. Regulating weapon autonomy through this multilateral forum is a particularly tough nut to crack. As I will argue in this section, this is due to two reasons. First, weapon autonomy as an issue is comparatively elusive and hard to conceptualize. Second, its perceived military value is exceptionally high, and the current geopolitical landscape is not conducive to new arms control breakthroughs.

Any discussion of the conceptual challenges regarding weapon autonomy has to begin with pointing out a common misunderstanding: the lack of progress in the CCW cannot be attributed to States Parties not having arrived at a shared definition of LAWS yet.[9] Quite to the contrary, it has much more to do with the fact that the attempt to define LAWS was misconceived from the very beginning. This warrants further elaboration.

The first two to three years of the CCW process on LAWS were indeed plagued by confusion and definitional struggles. Considerable effort was required to delineate the LAWS debate from the disputes surrounding remotely piloted aerial vehicles (drones) as well as to avoid anthropomorphizing LAWS as a one-to-one replacement for human soldiers.[10] All stakeholders were seeking – and quite a few lamenting the lack of – a "possible definition of LAWS", sometimes deliberately so in order to justify political heel-dragging. The underlying rationale was that arms control always requires a precise categorization of the *object* in question, such as a landmine, before any regulative action can be taken.

7 Future of Life Institute (FLI), "Autonomous Weapons: An Open Letter from AI and Robotics Researchers", 28 July 2015, available at: https://futureoflife.org/open-letter-autonomous-weapons/; FLI, "An Open Letter to the United Nations Convention on Certain Conventional Weapons", 21 August 2017, available at: https://futureoflife.org/autonomous-weapons-open-letter-2017/.

8 Mary Wareham, "As Killer Robots Loom, Demands Grow to Keep Humans in Control of Use of Force", Human Rights Watch, 2020, available at: www.hrw.org/world-report/2020/country-chapters/killer-robots-loom-in-2020.

9 The need to arrive at a shared definition of LAWS remains a common notion among the CCW States Parties, and some still view it as a prerequisite for the talks to go anywhere. As an example for this line of thought, see the chair's summary of the discussion of the 2019 GGE meeting: "Some delegations chose to address the issue of definitions, with several different views on the need for definitions – working or otherwise – to make further progress in the work of the Group." UN, *Report of the 2019 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems: Chair's Summary*, UN Doc. CCW/GGE.1/2019/3/Add.1, 8 November 2019, p. 3, available at: https://tinyurl.com/y68rzkub.

10 Léonard van Rompaey, "Shifting from Autonomous Weapons to Military Networks", *Journal of International Humanitarian Legal Studies*, Vol. 10, No. 1, 2019, pp. 112–119, available at: https://brill.com/view/journals/ihls/10/1/article-p111_111.xml.

https://doi.org/10.1017/S1816383120000466 Published online by Cambridge University Press

In the case of LAWS, however, the old pattern of defining and then regulating a discrete category of military hardware is not applicable.[11] After all, almost any current and future weapons system can conceivably be endowed with autonomous functions, and no one will be able to tell what any given system's level of dependence on human input is by merely inspecting it from the outside. In the past, bilateral nuclear arms control between the United States and the Soviet Union, later Russia, implemented quantitative arms control by developing precisely defined, shared understandings of counting rules and employing them in verification regimes.[12] Similarly, in the realm of multilateral conventional arms control, the now defunct Treaty on Conventional Armed Forces in Europe relied heavily on defining and counting military hardware items.[13] The challenge regarding LAWS, however, is not met by trying to define a category of weapons system – "LAWS", as separated with a list of specific criteria from "non-LAWS" – and then counting and capping its numbers. In fact, in a modern military's system-of-systems architecture, "some AWS [autonomous weapons system] components are intangible and can be geographically distributed, [so] it is far from clear … where and when an AWS begins and ends".[14] Hence, the challenge, broadly speaking, lies in developing a new norm in order to adjust the relationship between humans and machines in twenty-first-century warfighting. A qualitative rather than quantitative approach is required, which, in turn, requires new diplomatic language to grasp the underlying technological developments, something that neither States Parties nor civil society are well versed in yet.

Luckily, the process of conceptualizing the issue and translating it into diplomatic language has begun, and has already made some progress. After almost six years, the codification of a positive obligation of human control over weapons systems is establishing itself more and more at the heart of the debate. This general notion, gaining prominence in the wake of the call for "*meaningful* human control" originally introduced by the NGO Article 36,[15] is being embraced by civil society as well as a consistently growing number of CCW States Parties. Accordingly, the conceptualization is now finding broad acceptance in both academic literature and the diplomatic debate – not least because the United States and the International Committee of the Red Cross (ICRC) have adopted it. This is not some sort of categorical definition of LAWS (versus

11 Elvira Rosert and Frank Sauer, "How (Not) to Stop the Killer Robots: A Comparative Analysis of Humanitarian Disarmament Campaign Strategies", *Contemporary Security Policy*, 30 May 2020, available at: https://tinyurl.com/y23o8lo6.

12 Jozef Goldblat, *Arms Control: The New Guide to Negotiations and Agreements*, SAGE Publications, London, 2002, Chap. 5.

13 Treaty on Conventional Armed Forces in Europe, 19 November 1990, available at: www.osce.org/library/14087.

14 Maya Brehm, *Defending the Boundary: Constraints and Requirements on the Use of Autonomous Weapon Systems Under International Humanitarian and Human Rights Law*, Geneva Academy Briefing No. 9, May 2017, pp. 15–16.

15 Richard Moyes, "Key Elements of Meaningful Human Control", Article 36, April 2016, available at: www.article36.org/wp-content/uploads/2016/04/MHC-2016-FINAL.pdf. Article 36 is a member of the KRC.

https://doi.org/10.1017/S1816383120000466 Published online by Cambridge University Press

non-LAWS) via a list of criteria. Instead, it is a functional understanding of the phenomenon.[16]

From a functionalist point of view, the LAWS issue is best understood as one of autonomy *in* a weapons system – that is, of the machine rather than a human performing a certain function (or certain functions) during the system's operation.[17] Every military operation concluding with an attack on a target can be systematized along discrete steps of a kill chain or targeting cycle.[18] This includes finding, fixing, tracking, selecting and engaging the target (as well as assessing the effects afterwards). Many weapons systems are already capable of performing some of the targeting cycle functions without human input or supervision – for example, a drone navigating from one waypoint to the next via satellite navigation and thus performing a part of the "finding" function without having to be remotely controlled. An autonomous weapon, however, completes the entire targeting cycle – including the final stages of selecting and engaging the target with force – without human intervention. In the debate about LAWS, the focus rests mainly on those last two functions (which the ICRC calls "critical"[19]) because most of the effects of weapon autonomy currently under discussion derive from giving up human control over them and handing the decision to use force over to a machine.[20]

A peculiarity of the functional approach is that it reminds us that weapons with autonomy in their critical functions already exist. It renders the issue one of the present, not a concern about future weapons technology. That said, so far weapon autonomy, including the critical functions of target selection and engagement, exists only in limited military applications. The Israeli loitering munition Harpy is probably the best example of an already existing weapons system that – albeit only for the very specific task of engaging radar signatures – selects and engages targets without human supervision or control.[21] Harpy is thus considered an autonomous weapons system, as it completes a targeting cycle without human

16 ICRC, *Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons*, Geneva, 2016; US Department of Defense (DoD), Directive 3000.09, "Autonomy in Weapon Systems", 2012 (amended 2017); Paul Scharre, *Army of None: Autonomous Weapons and the Future of War*, W. W. Norton, New York, 2018.

17 Vincent Boulanin and Maaike Verbruggen, *Mapping the Development of Autonomy in Weapon Systems*, Stockholm International Peace Research Institute (SIPRI), Stockholm, 2017, available at: www.sipri.org/sites/default/files/2017-11/siprireport_mapping_the_development_of_autonomy_in_weapon_systems_1117_0.pdf.

18 International Panel on the Regulation of Autonomous Weapons (iPRAW), *Focus on Human Control*, iPRAW Report No. 5, August 2019, available at: www.ipraw.org/wp-content/uploads/2019/08/2019-08-09_iPRAW_HumanControl.pdf.

19 ICRC, above note 16, p. 7.

20 For the implications of autonomy in earlier stages of the targeting cycle, which are not discussed further here, see Arthur H. Michel, "The Killer Algorithms Nobody's Talking About", *Foreign Policy*, 20 January 2020, available at: https://foreignpolicy.com/2020/01/20/ai-autonomous-weapons-artificial-intelligence-the-killer-algorithms-nobodys-talking-about/.

21 Israel Aerospace Industries, "HARPY: Autonomous Weapon for All Weather", available at: *www.iai.co.il/p/harpy*. A loitering munition is a weapons system that "loiters" in an area for a prolonged period of time, waiting for targets to appear.

https://doi.org/10.1017/S1816383120000466 Published online by Cambridge University Press

intervention. Terminal defence systems capable of firing without human input, such as Phalanx or Patriot, are additional examples.

Especially in the earlier phases of the CCW process, systems like Phalanx gave rise to attempts by some States Parties to classify terminal defence systems as automatic, in order to prevent them from being drawn into the autonomy debate. In this line of reasoning, automatic systems are stationary and are designed to merely repeat a few pre-programmed actions in case of incoming munitions, whilst operating within tightly set parameters and time frames in structured and controlled environments. Autonomous systems, by contrast, are conceived of as having more time and operational room for manoeuvre.[22] Unfortunately, from an engineering point of view, no such clear and commonly agreed upon delineation between automaticity and autonomy exists; in fact, the terms "automatic" and "autonomous" can be, and often are, used interchangeably to describe a process in which a function is being performed by a machine rather than a human.[23]

A functional understanding renders any attempt at an automatic/autonomous delineation superfluous. This is an advantage in terms of conceptual clarity and simplicity. Also, what initially gave rise to the LAWS debate were concerns regarding autonomous targeting of humans, not targeting of missiles, mortar shells or other munitions.[24] Hence, the crux of the matter, also regarding possible regulation, is not whether a system is to be considered automatic or autonomous, but which targets it attacks. I will return to this line of thought in the sections on why regulation is imperative from an ethical point of view and how it is feasible.

Another advantage of the functionalist view is that it allows us to remain largely agnostic regarding the sophistication or the precise characteristics of the underlying technology. Terminal defence systems, to stick with them as an example, have been in use for decades. So, techniques such as machine learning (or whatever is currently *en vogue* in the wide field that is AI) are not necessarily required to lend a weapons system autonomy (or, for that matter, automaticity) in the critical functions of target selection and engagement. That said, AI obviously is a new and powerful enabler. Weapon autonomy is thus not really new, but recent innovations in AI, such as computer vision, are allowing actors to utilize weapon autonomy on a much larger scale. In effect, it is only recently that autonomous targeting has started to leave its former military niche applications and become adoptable across the board.

22 Frank Sauer, "Stopping 'Killer Robots': Why Now Is the Time to Ban Autonomous Weapons Systems", *Arms Control Today*, Vol. 46, No. 8, 2016, pp. 8–9.

23 To give but one example, the J3016 Levels of Automated Driving standard issued by the Society of Automotive Engineers (SAE) "defines six levels of driving automation" and considers level 5 to be "full vehicle autonomy". SAE, "SAE Standards News: J3016 Automated-Driving Graphic Update", 7 January 2019, available at: www.sae.org/news/2019/01/sae-updates-j3016-automated-driving-graphic.

24 I owe this point to an anonymous reviewer.

https://doi.org/10.1017/S1816383120000466 Published online by Cambridge University Press

Autonomous targeting being used only to destroy incoming munitions may feed into a general worry regarding the ever-faster pace of combat,[25] but it is of no humanitarian concern, and it helps to protect soldiers' lives – a fact that would have to be taken into consideration in any possible regulation of LAWS. In stark contrast, autonomy used unmitigatedly in all kinds of weapons systems, in various operational contexts, and against not just incoming munitions but *any and all targets*, including humans, creates more risks than benefits in sum, as will be argued in more detail below.

To sum up, the first reason why regulating weapon autonomy is difficult derives from the fact that CCW States Parties are challenged not to find some common definition of LAWS but instead to collectively stipulate how future targeting processes can be designed so that human control over the use of military force is retained.[26] In other words, the challenge lies not in delineating a specific weapons category but in generally regulating when a machine should make a certain decision or perform a certain function and when a human should do so, especially at the last two stages of the targeting cycle.

This endeavour is further complicated by the fact that, depending on the operational context and the nature of the target, the manner in which human control is implemented can vary. The combat direction system of a navy frigate, for instance, if designed only to fire at incoming anti-ship missiles and operated in autonomous mode only for brief periods of time whilst in the uncluttered environment of the sea, can be considered as remaining under human control "in design and use"[27] even while performing the critical functions of target selection and engagement autonomously. In contrast, an AI-enabled gun designed to accelerate targeting on a main battle tank in an urban environment would require every single shell fired to be triggered by a human with sufficient situational awareness to make an informed decision in order to be considered as remaining under human control in a meaningful sense.

In short, there is no one-size-fits-all standard of meaningful human control[28] because control by design requires a minimum standard of human–machine interaction, whereas control in use is implemented on a case-by-case

25 The general notion of an action–reaction dynamic created by increasing autonomy was first described by Jürgen Altmann: "Because of very fast action and reaction, autonomous weapon systems would create strong pressures for fast attack if both opponents have got them." Jürgen Altmann, "Military Uses of Nanotechnology: Perspectives and Concerns", *Security Dialogue*, Vol. 35, No. 1, 2004, p. 63.

26 Maya Brehm, "Targeting People", Article 36, November 2019, available at: *www.article36.org/wp-content/uploads/2019/11/targeting-people.pdf*; Richard Moyes, "Autonomy in Weapons Systems: Mapping a Structure for Regulation Through Specific Policy Questions", Article 36, November 2019, available at: www.article36.org/wp-content/uploads/2019/11/regulation-structure.pdf; Richard Moyes, "Target Profiles", Article 36, August 2019, available at: https://t.co/HZ1pvMnIks?amp=1; iPRAW, above note 18; Vincent Boulanin, Neil Davison, Netta Goussac and Moa Peldán Carlsson, *Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control*, SIPRI and ICRC, June 2020, available at: www.sipri.org/sites/default/files/2020-06/2006_limits_of_autonomy.pdf.

27 iPRAW, above note 18, pp. 12–13.

28 Daniele Amoroso and Guglielmo Tamburrini, *What Makes Human Control over Weapon Systems "Meaningful"?*, International Committee for Robot Arms Control, August 2019, available at: www.icrac.net/wp-content/uploads/2019/08/Amoroso-Tamburrini_Human-Control_ICRAC-WP4.pdf.

https://doi.org/10.1017/S1816383120000466 Published online by Cambridge University Press

basis.[29] And while defending against incoming munitions remains a worthwhile application of autonomy in a weapon's critical functions, the debate around LAWS suggests that arguably all other targets might require more human involvement and control. This renders the issue of LAWS more abstract, complex, and intellectually and diplomatically challenging than, for instance, conceptualizing a prohibition against anti-personnel landmines.

The second reason why regulating autonomy in weapons systems is difficult is the enormous military significance ascribed to it. This pertains to the five permanent members of the UN Security Council, but also to other countries with technologically advanced militaries such as, to give but two examples, Israel and Australia. The hurdle itself is not new, of course. It is observable in other regulatory processes of the recent past, such as the ones on landmines, cluster munitions and blinding laser weapons, with the latter being achieved within the CCW framework.[30] However, blinding lasers always represented an exotic niche capability that States could forego without great perceived military costs. Landmines and cluster munitions, too, had specific fields of use and were at least partly substitutable. This is not the case with weapon autonomy. Its impact is perceived to be game-changing for militaries in at least two domains of major significance.

First, weapon autonomy promises a whole range of operational and strategic advantages by rendering constant control and communication links obsolete. The militarily beneficial effects of this innovation, proponents argue, are manifold. It allows for a new level of force multiplication (with a single human operating several, dozens or hundreds of systems at once), creates the possibility of "swarming" (opening up new possibilities for overwhelming the enemy and evading counter-fire),[31] reduces personnel costs and increases a system's stealth in the electromagnetic spectrum (offering insurance against communications disruption or hijacking). Most importantly, however, it removes the inevitable delay between a remote human operator's command and the system's response. Swifter reaction times generate a key tactical advantage over a remotely controlled and thus slower-reacting adversarial system. In fact, the promise of gaining the upper hand by allowing for the completion of the targeting cycle at machine speed is arguably the main motivation behind increasing weapon autonomy.[32] Second, weapon autonomy promises to help prevent some of the atrocities of war and render warfare more humane. Since machines know no fear, stress or fatigue and are devoid of negative human emotions, they never panic, overreact or seek revenge, it is argued. Since they lack a self-preservation instinct,

29 I am thankful to an anonymous reviewer for this clarification.

30 E. Rosert and F. Sauer, above note 11.

31 Paul Scharre, *Robotics on the Battlefield, Part II: The Coming Swarm*, Center for a New American Security (CNAS), October 2014, available at: https://tinyurl.com/yy4gxs43; Maaike Verbruggen, *The Question of Swarms Control: Challenges to Ensuring Human Control over Military Swarms*, EU Non-Proliferation and Disarmament Consortium, Non-Proliferation and Disarmament Paper No. 65, December 2019.

32 Michael C. Horowitz, "When Speed Kills: Lethal Autonomous Weapon Systems, Deterrence and Stability", *Journal of Strategic Studies*, Vol. 42, No. 6, 2019; Jürgen Altmann and Frank Sauer, "Autonomous Weapon Systems and Strategic Stability", *Survival*, Vol. 59, No. 5, 2017.

https://doi.org/10.1017/S1816383120000466 Published online by Cambridge University Press

they can always delay returning fire. They supposedly allow not only for greater restraint but also – eventually, when technology permits – for better discrimination between civilians and combatants, thus resulting in the potential to apply military force in stricter accordance with the rules of international humanitarian law (IHL). This would add up to an overall ethical benefit – in a utilitarian sense.[33] In sum, the perceived transformative potential of weapon autonomy and the quantity and quality of military benefits ascribed to it render it more significant when compared to specific weapon categories, such as landmines or cluster munitions, that have been subject to humanitarian disarmament in the recent past.

In light of such tempting promises and the already ongoing, expanding efforts begun in the United States and China (as well as Russia, to a slightly lesser extent) to incorporate civilian innovation for the purposes of increasing weapon autonomy,[34] there is currently little appetite in those States to forego some of the conceived military benefits of this, in their view, critical step in military technology.[35] The United States' unwavering position, for instance, is to keep exploring an IHL-compliant use of autonomy in the critical functions of weapons systems.[36] Some middle powers are not keen on regulation at this point either. India, for example, senses an opportunity for leapfrogging and closing the technological gap between itself and the high-tech militaries of the world.[37] In fact, after the campaigns against landmines and cluster munitions, and the current humanitarian disarmament efforts in the areas of the arms trade (Arms Trade Treaty) and nuclear weapons (Treaty on the Prohibition of Nuclear Weapons), some diplomats in Geneva seem outright annoyed by the KRC's push for yet another prohibition treaty.

In addition, geopolitics in general are not conducive to achieving new arms control breakthroughs. The Treaty on the Prohibition of Nuclear Weapons, which will come into force on 22 January 2021, is seen by some as the exception to this rule,

33 Ronald C. Arkin, "Ethical Robots in Warfare", *IEEE Technology and Society Magazine*, Vol. 28, No. 1, 2009; Ronald C. Arkin, "The Case for Ethical Autonomy in Unmanned Systems", *Journal of Military Ethics*, Vol. 9, No. 4, 2010; Ronald C. Arkin, "Governing Lethal Behavior in Robots", *IEEE Technology and Society Magazine*, Vol. 30, No. 4, 2011; United States, *Implementing International Humanitarian Law in the Use of Autonomy in Weapon Systems: Working Paper Submitted by the United States of America*, UN Doc. CCW/GGE.1/2019/WP.5, 28 March 2019, available at: https://tinyurl.com/y4xe7tmc.

34 Elsa B. Kania, "In Military-Civil Fusion, China Is Learning Lessons from the United States and Starting to Innovate", *The Strategy Bridge*, 27 August 2019, available at: https://thestrategybridge.org/the-bridge/2019/8/27/in-military-civil-fusion-china-is-learning-lessons-from-the-united-states-and-starting-to-innovate; Elsa B. Kania, *"AI Weapons" in China's Military Innovation*, Brookings Institution, April 2020, available at: www.brookings.edu/wp-content/uploads/2020/04/FP_20200427_ai_weapons_kania_v2.pdf; Frank Sauer, "Military Applications of Artificial Intelligence: Nuclear Risk Redux", in Vincent Boulanin (ed.), *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*, SIPRI, Stockholm, 2019.

35 Michael C. Horowitz, "Artificial Intelligence, International Competition, and the Balance of Power", *Texas National Security Review*, Vol. 1, No. 3, 2018, available at: https://tnsr.org/2018/05/artificial-intelligence-international-competition-and-the-balance-of-power/; Zachary Davis, "Artificial Intelligence on the Battlefield: Implications for Deterrence and Surprise", *Prism*, Vol. 8, No. 2, 2019, pp. 117–121.

36 United States, above note 33. I would like to thank an anonymous reviewer for highlighting this.

37 Shashank R. Reddy, *India and the Challenge of Autonomous Weapons*, Carnegie Endowment for International Peace, June 2016, p. 12, available at: https://carnegieendowment.org/files/CEIP_CP275_Reddy_final.pdf.

https://doi.org/10.1017/S1816383120000466 Published online by Cambridge University Press

but its effects on the multilateral nuclear arms control architecture are still unclear. And at the same time, already existing multilateral and bilateral agreements and treaties are eroding, with some already lost – this list includes the terminated Intermediate-Range Nuclear Forces Treaty, the faltering Joint Comprehensive Plan of Action with Iran, the struggling Open Skies Treaty, and soon, potentially, NewSTART, the only remaining bilateral nuclear arms control treaty between Russia and the United States. Getting a new binding international legal instrument out of the CCW would be challenging in a normal, less frosty geopolitical landscape. The current global arms control winter makes it a daunting feat.

Nevertheless, regulating weapon autonomy in a manner that curbs autonomy in the critical functions and keeps them under human control is sorely needed. After all, the consequences of inaction would be dire because the mid- and long-term strategic and ethical risks of unshackled weapon autonomy far outweigh the desired short-term military gains highlighted above. I will argue this in two steps below, by first focusing on a number of operational and strategic implications and subsequently evaluating the ethical implications of weapon autonomy in regard to human dignity.

Why regulating weapon autonomy is imperative: Strategic implications

The potential impact of unregulated weapon autonomy on military operations, as well as on global peace and strategic stability as a whole, has drawn scholarly attention for quite a while.[38] This body of literature suggests that the implications of regulatory inaction and an ensuing rapid diffusion of weaponized autonomy-enabling technology range from new military vulnerabilities to increased risks of instability and escalation at both the operational and the strategic level.[39] Hence it is in fact especially the great powers that should see it as being not only their responsibility but also in their genuine self-interest[40] to curb this destabilizing chain of effects.

38 See J. Altmann, above note 25; Armin Krishnan, *Killer Robots: Legality and Ethicality of Autonomous Weapons*, Ashgate, Farnham, 2009, Chap. 6; Jean-Marc Rickli, *Some Considerations of the Impact of LAWS on International Security: Strategic Stability, Non-State Actors and Future Prospects*, presentation at CCW Meeting of Experts on LAWS, Geneva, 16 April 2015, available at: https://tinyurl.com/y4fjozpf; Paul Scharre, *Autonomous Weapons and Operational Risk*, CNAS Ethical Autonomy Project, Washington, DC, February 2016, available at: https://s3.amazonaws.com/files.cnas.org/documents/CNAS_Autonomous-weapons-operational-risk.pdf; Wendell Wallach, "Toward a Ban on Lethal Autonomous Weapons: Surmounting the Obstacles", *Communications of the ACM*, Vol. 60, No. 5, 2017, p. 31; Irving Lachow, "The Upside and Downside of Swarming Drones", *Bulletin of the Atomic Scientists*, Vol. 73, No. 2, 2017; J. Altmann and F. Sauer, above note 32; Paul Scharre, "Autonomous Weapons and Stability", PhD thesis, King's College London, March 2020, available at: https://kclpure.kcl.ac.uk/portal/files/129451536/2020_Scharre_Paul_1575997_ethesis.pdf.

39 The following section draws on J. Altmann and F. Sauer, above note 32; F. Sauer, above note 34; Aaron Hansen and Frank Sauer, "Autonomie in Waffensystemen: Chancen und Risiken Für die US-Sicherheitspolitik", *Zeitschrift für Außen- und Sicherheitspolitik*, Vol. 12, No. 2, 2019.

40 For the general argument, see Hedley Bull, *The Anarchical Society: A Study of Order in World Politics*, Macmillan, London, 1977. For the case of AI, see Elsa B. Kania and Andrew Imbrie, "Great Powers

https://doi.org/10.1017/S1816383120000466 Published online by Cambridge University Press

Technology diffusion

To get an idea of the expectable diffusion of technology in the field of LAWS, drones can serve as an indicator.[41] China in particular is not only investing but also exporting in this sector.[42] According to data collected by the New America Foundation,[43] twelve countries have conducted drone strikes and thirty-eight now possess armed drones, as do several non-State actors such as Hamas, Hezbollah, the Houthi rebels and the so-called Islamic State group.

Drone technology spreads comparably quickly because of its dual-use nature. Autonomy is dual-use, too. Weapon autonomy – provided that the platform in question contains the necessary sensors and actuators – mainly comes down to software, which can be transferred and reproduced at close to no cost and is vulnerable to theft via cyber attacks.[44] Consequently, the adoption of software-enabled autonomous functions can be expected to spread rapidly in the existing military hardware ecosystem. Also, the main innovators of autonomy are tech companies and universities, not the defence industry, so it is questionable whether any one country's military can remain the "fast leader"[45] as envisioned by members of the US defence establishment, for example. After all, the US government is not the only one incorporating civilian tech for military purposes by approaching tech firms such as Google, Microsoft and Amazon; China, for example, is doing the same with Tencent, Ali Baba and Baidu.[46] Hence it is highly unlikely that some sort of monopoly in this field, like the one the United States held with stealth technology in the past, is possible.

New operational vulnerabilities

The flip side of the force multiplication effect that militaries hope for with this diffusion-prone technology is scalability, creating the potential for weaker parties to change the power dynamics between themselves and their adversaries. The

Must Talk to Each Other about AI", *Defense One*, 28 January 2020, available at: www.defenseone.com/ideas/2020/01/great-powers-must-talk-each-other-about-ai/162686/?oref=d-river.

41 Frank Sauer and Niklas Schörnig, "Killer Drones: The Silver Bullet of Democratic Warfare?", *Security Dialogue*, Vol. 43, No. 4, 2012; Matthew Fuhrmann and Michael C. Horowitz, "Droning On: Explaining the Proliferation of Unmanned Aerial Vehicles", *International Organization*, Vol. 71, No. 2, 2017; Andrea Gilli and Mauro Gilli, "The Diffusion of Drone Warfare? Industrial, Organizational, and Infrastructural Constraints", *Security Studies*, Vol. 25, No. 1, 2016.

42 Defense Science Board, *The Role of Autonomy in DoD Systems*, 2012, pp. 69–71.

43 New America, "World of Drones", available at: www.newamerica.org/in-depth/world-of-drones/.

44 Sydney J. Friedberg, "Robot Wars: Centaurs, Skynet, and Swarms", *Breaking Defense*, 31 December 2015, available at: http://breakingdefense.com/2015/12/robot-wars-centaurs-skynet-swarms/; Thomas G. Mahnken, *Technology and the American Way of War Since 1945*, Columbia University Press, New York, 2008, p. 123.

45 Robert O. Work, "Robert Work Talks NATO's Technological Innovation and the DoD", *CNAS Brussels Sprouts Podcast*, 11 January 2018, available at: www.cnas.org/publications/podcast/robert-work-talks-natos-technological-innovation-and-the-dod.

46 Defense Science Board, *Summer Study on Autonomy*, 2016, p. 45; Elsa B. Kania, *Battlefield Singularity: Artificial Intelligence, Military Revolution, and China's Future Military Power*, CNAS, Washington, DC, November 2017, available at: https://s3.amazonaws.com/files.cnas.org/documents/Battlefield-Singularity-November-2017.pdf?mtime=20171129235805; E. B. Kania, "In Military-Civil Fusion", above note 34.

https://doi.org/10.1017/S1816383120000466 Published online by Cambridge University Press

weaponization of simple, commercially available drones by the so-called Islamic State group and the attack against the Saudi Aramco oil facility give non-autonomous foretastes of what is to come and demonstrate that, advanced aerial defence capabilities notwithstanding, new vulnerabilities are on the rise. Particularly from the point of view of US ground forces, having to face serious threats from above after decades of air dominance represents a paradigm shift.[47] The United States is thus already being forced to rethink its air defence capabilities by intensifying the development of lasers and microwaves. Conventional solutions, such as Stinger missiles, are not only unsuitable for defence against the swarms of small, cheap, disposable drones that autonomy now renders possible, they are also not cost-effective. Whether the new defensive systems can remedy this situation is still open for debate.[48] Suffice it to say that the combination of cheap unmanned systems, autonomy and swarm behaviour creates new risks in general, for troops on the battlefield, for command and control infrastructure and for senior leaders in so-called decapitation scenarios.[49]

As argued above, the possible elimination of the remote control link is a key incentive for having more autonomy in a weapons system – but handing control over to the machine opens up new attack vectors as well. Feeding the system spoofed GPS data is one example; in 2011, Iran was seemingly able to hijack an autonomously navigating US drone in this manner.[50]

What is more, systems relying on machine learning that makes use of deep neural networks,[51] which currently represent the state of the art in the field of computer vision, are also particularly susceptible to manipulation. Some reflective tape on a stop sign, for example, can fool a self-driving car's image recognition system. This susceptibility to error is a tricky but eventually solvable problem in a civilian application such as self-driving cars. Training data is plentiful and easily available, and self-driving cars are designed to operate cooperatively in a tightly regulated environment. The battlefield presents itself very differently – it is characterized by a paucity of data and much greater degrees of unpredictability and vulnerability.[52] After all, an adversary will of course always try to deceive and

47 Kelley Sayler, *A World of Proliferated Drones: A Technology Primer*, CNAS, Washington, DC, 2015, p. 29.

48 Sebastien Roblin, "The U.S. Army Needs More Anti-Aircraft Weapons – and Fast", *War is Boring*, 22 January 2018, available at: http://warisboring.com/the-u-s-army-needs-more-anti-aircraft-weapons-and-fast/.

49 David Barno and Nora Bensahel, "The Drone Beats of War: The U.S. Vulnerability to Targeted Killings", *War on the Rocks*, 21 January 2020, available at: https://warontherocks.com/2020/01/the-drone-beats-of-war-the-u-s-vulnerability-to-targeted-killings/. A decapitation scenario is a scenario in which an attacker aims to destroy or destabilize an opponent's leadership and command and control structure in order to severely degrade or destroy its capacity for (nuclear) retaliation.

50 Sydney J. Friedberg, "Drones Need Secure Datalinks to Survive vs. Iran, China", *Breaking Defense*, 10 August 2012, available at: http://breakingdefense.com/2012/08/drones-need-secure-datalinks-to-survive-vs-iran-china/.

51 For a critical overview, see Gary Marcus, "Deep Learning: A Critical Appraisal", New York University, 2 January 2018, available at: https://arxiv.org/ftp/arxiv/papers/1801/1801.00631.pdf.

52 Michał Klincewicz, "Autonomous Weapons Systems, the Frame Problem and Computer Security", *Journal of Military Ethics*, Vol. 14, No. 2, 2015; Anh Nguyen, Jason Yosinski and Jeff Clune, "Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images",

https://doi.org/10.1017/S1816383120000466 Published online by Cambridge University Press

tamper with your systems. Research on adversarial examples[53] suggests that computer vision will leave autonomous weapons systems open to manipulation by tampering with the environment that the machines perceive[54] or even by retraining them if they continue learning during their deployment.[55] Facial recognition for targeting purposes would be quite easy to fool and defeat, too, as the rapid development of countermeasures against domestic surveillance demonstrates.[56]

As the complexity of the software driving a weapons system increases, so does the number of bugs it contains. Such programming errors can have critical effects, including friendly fire.[57] Normal accidents theory[58] suggests that mistakes are basically inevitable. They occur even in domains with extremely high safety and security standards, such as nuclear power plants or manned space travel.[59] The software industry can currently reduce the number of bugs to 0.1–0.5 errors per 1,000 lines of code, which means that complex military systems with several million lines of code, such as the software for the F-35 fighter jet, come with thousands of software errors.[60] The unavoidable reality of regularly having to update the systems complicates this issue further;[61] it is a potential source of new bugs and new errors arising from interactions between newer and older software. Machine learning systems generate specific difficulties because they present themselves as "black boxes" which cannot be debugged the way conventional software can, meaning that they cannot be selectively cleared of specific errors.[62]

Finally, weapon autonomy evokes a new proneness to errors in regard to any remaining interactions with human operators. Here, automation bias comes into play – that is, the uncritical, unfounded trust in the functioning of a

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 427–436; Z. Davis, above note 35, pp. 121–122.

53 Ivan Evtimov *et al.*, "Robust Physical-World Attacks on Deep Learning Models", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, available at: https://arxiv.org/pdf/1707.08945.pdf.

54 See the by now famous example of the turtle mistaken for a rifle, in Anish Athalye, Logan Engstrom, Andrew Ilyas and Kevin Kwok, "Synthesizing Robust Adversarial Examples", *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80, 2018, available at: https://arxiv.org/pdf/1707.07397.pdf.

55 Defense Science Board, above note 46, p. 28; Vadim Kozyulin, "International and Regional Threats Posed by the LAWS: Russian Perspective", PIR Center for Policy Studies, April 2016, available at: https://tinyurl.com/y4qslefc; P. Scharre, *Autonomous Weapons and Operational Risk*, above note 38, p. 14.

56 Melissa Hellmann, "Special Sunglasses, License-Plate Dresses: How to Be Anonymous in the Age of Surveillance", *Seattle Times*, 12 January 2020, available at: www.seattletimes.com/business/technology/special-sunglasses-license-plate-dresses-juggalo-face-paint-how-to-be-anonymous-in-the-age-of-surveillance/.

57 P. Scharre, above note 38, p. 21.

58 Charles Perrow, *Normal Accidents: Living with High-Risk Technologies*, Basic Books, New York, 1984.

59 John Borrie, *Security, Unintentional Risk, and System Accidents*, United Nations Institute for Disarmament Research (UNIDIR), Geneva, 15 April 2016, available at: https://tinyurl.com/yyaugayk; P. Scharre, *Autonomous Weapons and Operational Risk*, above note 38.

60 P. Scharre, *Autonomous Weapons and Operational Risk*, above note 38, p. 13.

61 UNIDIR, *The Weaponization of Increasingly Autonomous Technologies in the Maritime Environment: Testing the Waters*, UNIDIR Resources No. 4, Geneva, 2015, p. 8.

62 G. Marcus, above note 51, pp. 10–11.

https://doi.org/10.1017/S1816383120000466 Published online by Cambridge University Press

system.[63] Put simply, autonomous systems might operate incorrectly over periods of time without anyone noticing.[64] A human making a mistake can understand the situation and correct for it, but unmonitored LAWS will not be able to understand and critically reflect in real time the way humans do.[65] This gives rise to risks of unintended military escalation.

Escalation risks and crisis instability

Weapons systems operating without human control generate not only new vulnerabilities but also unpredictability due to unforeseeable interactions with their environment, in turn creating new risks of unintended, unwanted escalation.[66] In that regard, the interaction between two or more autonomous systems is to be considered in particular. High-frequency trading[67] provides a useful analogue, because unforeseen and unwanted interaction processes between two or more autonomously operating trading algorithms occur on a regular basis, sometimes causing so-called "flash crashes" and resulting in financial losses. This can be remedied with regulation of the financial market to an extent, but without internationally binding regulation of autonomy on the battlefield, unforeseeable interactions of LAWS might end in unintended use of force at machine speed, even accidental war before humans can intervene.[68] This risk is not in some distant future. At the Dubai Airshow in 2019, the chief of staff of the US Air Force, General David Goldfein, presented the simulated engagement of an enemy navy vessel with a next-to-fully automated kill chain. The vessel was first picked up by a satellite, then target data was relayed to airborne surveillance as well as command and control assets. A US Navy destroyer was then tasked with firing a missile, the only remaining point at which this targeting cycle involved a human decision, with the rest of the "kill chain … completed machine to machine, at the speed of light".[69] Any machine error in such a system would, if left uncorrected

63 See, for the example of the Patriot missile defence system, John K. Hawley, *Patriot Wars: Automation and the Patriot Air and Missile Defense System*, CNAS, Washington, DC, January 2017, available at: www.cnas.org/publications/reports/patriot-wars.

64 P. Scharre, *Autonomous Weapons and Operational Risk*, above note 38, p. 31; Noel Sharkey and Lucy Suchman, "Wishful Mnemonics and Autonomous Killing Machines", *Proceedings of the AISB*, Vol. 136, 2013, pp. 16–17.

65 Defense Science Board, above note 42, p. 15.

66 André Haider and Maria Beatrice Catarrasi, *Future Unmanned System Technologies: Legal and Ethical Implications of Increasing Automation*, Joint Air Power Competence Centre, November 2016, p. 10, available at: www.japcc.org/wp-content/uploads/Future_Unmanned_System_Technologies_Web.pdf; ICRC, *Views of the International Committee of the Red Cross (ICRC) on Autonomous Weapon System [s]*, Geneva, 11 April 2016, p. 3, available at: www.icrc.org/en/download/file/21606/ccw-autonomous-weapons-icrc-april-2016.pdf.

67 Gary Shorter and Rena S. Miller, *High-Frequency Trading: Background, Concerns, and Regulatory Developments*, Congressional Research Service, 19 June 2014, available at: *https://fas.org/sgp/crs/misc/R43608.pdf*.

68 P. Scharre, *Autonomous Weapons and Operational Risk*, above note 38, p. 53; J. Altmann and F. Sauer, above note 32, pp. 128–132.

69 "Video: Here's How the US Air Force Is Automating the Future Kill Chain", *Defense News*, 2019, available at: www.defensenews.com/video/2019/11/16/heres-how-the-us-air-force-is-automating-the-future-kill-chain-dubai-airshow-2019/.

https://doi.org/10.1017/S1816383120000466 Published online by Cambridge University Press

by a human due to automation bias, propagate quickly. It stands to reason that the error would propagate "at the speed of light" as well, were the human to be removed. A recent wargaming exercise conducted by the RAND Corporation underlines the risks of crisis instability and unintended escalation; in this exercise, simulated forces were set "on 'full auto' to signal resolve …[,] in one case lead[ing] to inadvertent escalation. Systems set to autonomous mode reacted with force to an unanticipated situation in which the humans did not intend to use force."[70]

Humans are more resistant to mass error than machines. Also, humans, despite being slower and sometimes making mistakes, are better managers than machines. They have the capacity to grasp an unusual situation and understand its context as well as to reflect on a decision, its genesis, its implications and the weight of the responsibility that accompanies it. In terms of crisis management, all this makes humans superior to machines, which so far are only capable of recognizing patterns and executing predefined actions, and which reach superhuman performance only in those narrowly defined scenarios for which they were specifically trained. By removing human control, the distinct role of humans as a versatile fail-safe mechanism is lost.

The prominent case of Lieutenant Colonel Stanislav Petrov renders this evident. The 1983 NATO exercise Able Archer was misunderstood by the Soviets as a cover for an attack with tactical nuclear forces. During this time, a Soviet early-warning satellite registered first one, then a couple more US nuclear intercontinental ballistic missile launches. Petrov, the watch officer in charge at the time, decided (correctly) that this had to be a false alarm and gave the all-clear up the chain of command, thus preventing further, potentially nuclear escalation in this tense situation. Petrov's decision could not have been made by a completely automated system. He later testified that he had arrived at his decision by following a gut feeling, by wondering about the nature of the supposed strike, and by drawing on his past experiences with the early-warning system that he deemed not fully trustworthy.[71] If the human on the destroyer in the next-to-fully automated kill chain presented by General Goldfein were ever to be removed, fully actualizing the key advantage of weapon autonomy that is fighting at machine speed, the "Petrov effect" would be lost. While, in that conventional scenario, this would not mean the inadvertent use of nuclear

70 Yuna H. Wong *et al.*, *Deterrence in the Age of Thinking Machines*, RAND Corporation, 2020, p. xi, available at: www.rand.org/pubs/research_reports/RR2797.html.

71 Bruce G. Blair, *The Logic of Accidental Nuclear War*, Brookings Institution, Washington, DC, 1993, p. 181; Richard Rhodes, *Arsenals of Folly: The Making of the Nuclear Arms Race*, Simon & Schuster, London, 2008, pp. 165–166; David E. Hoffman, *The Dead Hand: The Untold Story of the Cold War Arms Race and Its Dangerous Legacy*, Doubleday, New York, 2009, pp. 6–11, 94–95; Mark Gubrud, "Stopping Killer Robots", *Bulletin of the Atomic Scientists*, Vol. 70, No. 1, 2014; Michael C. Horowitz, Paul Scharre and Alexander Velez-Green, *A Stable Nuclear Future? The Impact of Autonomous Systems and Artificial Intelligence*, working paper, December 2019, pp. 13–14, available at: https://arxiv.org/ftp/arxiv/papers/1912/1912.05291.pdf; Paul Scharre, "A Million Mistakes a Second", *Foreign Policy*, 12 September 2018, available at: https://foreignpolicy.com/2018/09/12/a-million-mistakes-a-second-future-of-war/.

https://doi.org/10.1017/S1816383120000466 Published online by Cambridge University Press

weapons, strategic stability is nevertheless already being affected by the effort to increase autonomy in military systems.

Strategic instability

It has recently been suggested that AI as a decision-making aid to humans might help improve the performance of nuclear early-warning and command and control systems, thus reducing the risk of false alarms and inadvertent nuclear use.[72] That said, calls for complete automation in the nuclear realm – that is, for handing over the decision to use nuclear weapons from humans to machines – are practically non-existent.[73] But even with the proverbial push of the button not yet delegated to algorithms, the rush to increase autonomy in military applications and to automatize military processes increases the risk of nuclear stability.[74]

For instance, the increasing capacities of conventional weapons systems – including weapon autonomy – are beginning to affect the strategic level. This development has been described as the increasing "entanglement" of the nuclear and the conventional realm resulting, for example, from "non-nuclear threats to nuclear weapons and their associated command, control, communication, and information (C3I) systems".[75] Simply put, advanced conventional capabilities increasingly allow for nuclear assets to be put at risk. Autonomy in conventional weapons systems is one such advanced capability, thus feeding into this increasing entanglement and, in turn, deteriorating strategic stability.

One specific illustration of this dynamic is the deployment of stealthy unmanned aerial vehicles and the use of "swarming". Perdix is a swarming test program pursued by the US Air Force. In the future, drone swarms of this type might facilitate the search for dispersed mobile missile launchers. Another example is the use of maritime autonomous systems for hunting nuclear-powered ballistic missile submarines, known as SSBNs. The DARPA-funded Anti-Submarine Warfare Continuous Trail Unmanned Vessel is a program that resulted in the development of an autonomous trimaran called Sea Hunter, which

72 M. Horowitz, P. Scharre and A. Velez-Green, above note 71, p. 14; Philip Reiner and Alexa Wehsner, "The Real Value of Artificial Intelligence in Nuclear Command and Control", *War on the Rocks*, 4 November 2019, available at: https://warontherocks.com/2019/11/the-real-value-of-artificial-intelligence-in-nuclear-command-and-control/. On the resulting cyber vulnerabilities, see James Johnson, "The AI–Cyber Nexus: Implications for Military Escalation, Deterrence and Strategic Stability", *Journal of Cyber Policy*, Vol. 4, No. 3, 2019.

73 With the exception of Adam Lowther and Curtis McGiffin, "America Needs a 'Dead Hand'", *War on the Rocks*, 16 August 2019, available at: https://warontherocks.com/2019/08/america-needs-a-dead-hand/.

74 Edward Geist and Andrew J. Lohn, *How Might Artificial Intelligence Affect the Risk of Nuclear War?*, RAND Corporation, 2018, available at: www.rand.org/content/dam/rand/pubs/perspectives/PE200/PE296/RAND_PE296.pdf; Vincent Boulanin, Lora Saalman, Petr Topychkanov, Fei Su and Moa Peldán Carlsson, *Artificial Intelligence, Strategic Stability and Nuclear Risk*, SIPRI, Stockholm, June 2020, available at: www.sipri.org/sites/default/files/2020-06/artificial_intelligence_strategic_stability_and_nuclear_risk.pdf.

75 James M. Acton (ed.), *Entanglement: Chinese and Russian Perspectives on Non-Nuclear Weapons and Nuclear Risks*, Carnegie Endowment for International Peace, 2017, p. 1, available at: http://carnegieendowment.org/files/Entanglement_interior_FNL.pdf.

https://doi.org/10.1017/S1816383120000466 Published online by Cambridge University Press

is currently being tested by the US Navy. Its ability to detect and pursue SSBNs could potentially limit the second-strike capabilities of other nuclear powers.

These capabilities are just emerging, and neither Perdix nor Sea Hunter, nor their successors, will single-handedly destabilize the global nuclear order. Also, the hypothesis that systems such as Sea Hunter would render the oceans "transparent",[76] virtually nullifying the utility of sea-launched nuclear weapons as a reliable second-strike asset, is hotly debated. Nevertheless, the mere perception of nuclear capabilities becoming susceptible to new risks from the conventional realm is bound to sow distrust between nuclear-armed adversaries. Furthermore, a system like Sea Hunter demonstrates how autonomous weapon technologies are expediting the completion of the targeting cycle, thus putting the adversary under additional pressure and potentially creating "use-them-or-lose-them" scenarios with regard to executing a nuclear second strike.

The entanglement problem, which weapon autonomy is feeding into, is further aggravated by an increasing political willingness to use nuclear means to retaliate against non-nuclear attacks on early-warning and control systems or the weapons themselves. The Trump administration's nuclear posture review[77] signals that the United States may, from now on, respond with nuclear means to significant, non-nuclear strategic attacks (moving away from a "single-purpose" nuclear deterrence framing for nuclear weapons). Russia has already held this position for some time due to the United States' advantage in conventional weapons technology. This does not bode well for stability between the two largest nuclear powers.

To sum up this section, weapon autonomy not only promises military benefits but also creates new vulnerabilities and, more importantly, contributes to an overall accumulation of strategic risk and instability. Increasing operational speed beyond the capability of human cognition removes humans as a valuable fail-safe against unwanted escalation.

Why regulating weapon autonomy is imperative: Ethical implications

The discussion on LAWS in the CCW is slanted towards IHL and the legal implications, as visible, for example, in the eleven guiding principles adopted in the 2019 CCW States Parties meeting report.[78] In the preamble preceding the list of principles, the report states that "international law, in particular the United Nations Charter and International Humanitarian Law … as well as relevant ethical perspectives, should guide the continued work of the Group". Nevertheless, only five out of the eleven guiding principles are legal in nature,

76 Sebastian Brixey-Williams, "Will the Atlantic Become Transparent?", November 2016, available at: https://britishpugwash.org/wp-content/uploads/2016/11/Will-the-Atlantic-become-transparent-.pdf.
77 DoD, *Nuclear Posture Review 2018*, 2018, p. 21, available at: https://tinyurl.com/yc7lu944.
78 CCW Meeting Final Report, above note 2, p. 10.

https://doi.org/10.1017/S1816383120000466 Published online by Cambridge University Press

and not a single one contains a reference to ethical implications. The legal strand of the debate is undoubtedly important, especially since it allows for systematically interrogating the claim that autonomy in weapons renders warfare more IHL-compliant. As it stands, technology is now unable to fulfil this promise of increased IHL compliance,[79] though it might eventually be capable of doing so. But be that as it may, an ethical point of view suggests that the LAWS issue has deeper roots than mere IHL compliance anyway, because it touches upon fundamental norms that go above and beyond the laws of war.[80] Ethical implications were more systematically considered in their own right at the very beginning of the LAWS debate at the UN. In 2013, when the issue was raised in the Human Rights Council, Special Rapporteur Christof Heyns[81] objected against LAWS, arguing that they violate human dignity.

Universal human dignity

That the use of LAWS would be a violation of human dignity has been argued by various scholars of moral philosophy and technology.[82] The notion was picked up by the KRC[83] and lately has also been reiterated by the ICRC.[84] Opposing weapon autonomy on grounds of human dignity has drawn some scrutiny,[85] and the supposed "awkwardness"[86] of this stance is commonly substantiated by pointing out that several meanings of dignity exist and that there is no commonly agreed-upon definition of dignity.

However, being hard to define but relevant and even crucially important is a characteristic of many normative concepts, including many legally codified ones. Cornerstones of IHL such as civilian-ness, which is defined only *ex negativo*, or proportionality, which is not quantifiable and is assessable only on a case-by-case

79 Frank Sauer, Daniele Amoroso, Noel Sharkey, Lucy Suchman and Guglielmo Tamburrini, *Autonomy in Weapon Systems: The Military Application of Artificial Intelligence as a Litmus Test for Germany's New Foreign and Security Policy*, Heinrich Böll Foundation Publication Series on Democracy, Vol. 49, 2018, pp. 23–32, available at: www.boell.de/sites/default/files/boell_autonomy-in-weapon-systems_v04_kommentierbar_1.pdf.

80 The following section draws on Elvira Rosert and Frank Sauer, "Prohibiting Autonomous Weapons: Put Human Dignity First", *Global Policy*, Vol. 10, No. 3, 2019.

81 Christof Heyns, *Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions*, UN Doc. A/HRC/23/47, 2013, p. 17, available at: https://digitallibrary.un.org/record/755741/files/A_HRC_23_47-EN.pdf.

82 Peter Asaro, "On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making", *International Review of the Red Cross*, Vol. 94, No. 886, 2012; Robert Sparrow, "Robots and Respect: Assessing the Case Against Autonomous Weapon Systems", *Ethics & International Affairs*, Vol. 30, No. 1, 2016.

83 KRC, *Making the Case: The Dangers of Killer Robots and the Need for a Preemptive Ban*, 2016, pp. 21–25.

84 ICRC, *Ethics and Autonomous Weapon Systems: An Ethical Basis for Human Control?*, Geneva, 3 April 2018, available at: www.icrc.org/en/download/file/69961/icrc_ethics_and_autonomous_weapon_systems_report_3_april_2018.pdf.

85 Amanda Sharkey, "Autonomous Weapons Systems, Killer Robots and Human Dignity", *Ethics and Information Technology*, Vol. 21, No. 2, 2019.

86 Deane-Peter Baker, "The Awkwardness of the Dignity Objection to Autonomous Weapons", *The Strategy Bridge*, 6 December 2018, available at: https://thestrategybridge.org/the-bridge/2018/12/6/the-awkwardness-of-the-dignity-objection-to-autonomous-weapons.

https://doi.org/10.1017/S1816383120000466 Published online by Cambridge University Press

basis, are examples.[87] Human dignity, too, is contained in various international legal documents. The Universal Declaration of Human Rights refers to it in its preamble, as does the UN Charter. It is also invoked in national bodies of law, as well as court decisions. The key example here is Germany's basic law Article 1(1), which states human dignity's inviolability and prohibits the treatment of humans as objects or means to an end, being referenced in a 2006 landmark decision by the German Constitutional Court. The judges struck down a federal law that would have allowed the German air force to shoot down a hijacked aeroplane that the hijackers may have intended to use as a weapon to kill people on the ground. The Court deemed it unconstitutional to use the aeroplane passengers as mere instruments to try to achieve another, albeit worthy, goal.[88]

The key ethical implication of weapon autonomy in a weapons system's critical functions is thus that allowing algorithms to make kill decisions violates human dignity because the victim is reduced to an object, a mere data point fed to an automated, indifferent killing machine.

It is worth spelling out that this objection is valid even if civilians (or other non-combatants) remain unharmed. After all, narrowing the focus solely to the possibility that LAWS might not be able to make proper – or even better – distinctions between combatants and civilians, a cornerstone of the legal case against LAWS discussed in the CCW, loses sight of the fact that combatants, too, are imbued with human dignity. In other words, weapon autonomy raises a more fundamental concern than the legal strand of the LAWS debate suggests, because "successfully discerning combatants from noncombatants is far from the only issue".[89]

As a general rule, the use of LAWS against humans can be deemed an unacceptable infringement on human dignity because delegating the decision to kill to an algorithm devalues human life.[90] Exceptions from this rule would only be conceivable if they were explicitly not made on the basis of weighing bare lives against each other and then deliberately opting for algorithmic killing. An example for such a boundary case could be a sailor's reliance on weapon autonomy in a narrowly bound scenario of desperate self-defence. If the aforementioned navy frigate[91] were to be under a saturation attack by anti-ship missiles and, potentially, also manned aircraft, then inadvertently endangering human life by relying on autonomous defensive fire for the survival of the ship and its crew could be considered acceptable *ex post*.

Generally speaking again, being killed as the result of algorithmic decision-making matters for the person dying because a machine taking a human life has no

87 I would like to thank an anonymous reviewer for specifying these properties of civilian-ness and proportionality.
88 F. Sauer *et al.*, above note 79, p. 33.
89 Heather M. Roff, "The Strategic Robot Problem: Lethal Autonomous Weapons in War", *Journal of Military Ethics*, Vol. 13, No. 3, 2014, p. 219.
90 Christof Heyns, "Autonomous Weapons in Armed Conflict and the Right to a Dignified Life: An African Perspective", *South African Journal on Human Rights*, Vol. 33, No. 1, 2017, pp. 62–63.
91 See text related to above note 27.

https://doi.org/10.1017/S1816383120000466 Published online by Cambridge University Press

conception of what its action means: "In the absence of an intentional and meaningful decision to use violence, the resulting deaths are meaningless and arbitrary."[92] In other words, the least we can do when killing another human being in war is to recognize this death of a fellow member of our species and put the weight accompanying this decision onto our conscience. The mindlessness of machines killing humans based on software outputs strips the latter of their right to be recognized as humans in the moment of death. This also matters for society at large. Modern warfare, especially in democracies, already decouples societies from warfighting in terms of political and financial costs.[93] A society outsourcing moral costs by no longer even concerning itself with the act of killing, with no individual combatants' psyches burdened by the accompanying responsibility, crosses a moral line. It risks losing touch with fundamental humanitarian values such as the right to a dignified life and respect towards fellow human beings.[94]

To sum up this section, while the legal verdict on weapon autonomy increasing IHL compliance is still out and will be for some time, a more fundamental objection against LAWS based on deontological limits is valid today.

How regulating weapon autonomy is feasible: Fostering a human control norm

The United States and China are demonstrating awareness of the strategic risks of unmitigated weapon autonomy. The US directive on weapon autonomy,[95] albeit attempting to square the circle of using autonomy while not inviting the accompanying risks, can be interpreted this way. China has coined the term "battlefield singularity", a dreaded situation in which war waged at machine speed is too fast for human cognition to keep up.[96] Nevertheless, the current great power rivalry between the United States, China and Russia, all racing for dominance in the field of military AI, is clearly not conducive to regulation of weapon autonomy. With presidents Trump, Xi and Putin in power, a breakthrough is not to be expected any time soon. But political will for regulation can also be generated from the ground up.

Growing political will from the grassroots

Surveys consistently show publics from all over the world rejecting LAWS. Opposition globally increased from 56% in 2016 to 61% in 2018, according to

92 Peter Asaro, "*Jus Nascendi*, Robotic Weapons, and the Martens Clause", in Ryan Calo, A. Michael Froomkin and Ian Kerr (eds), *Robot Law*, Edward Elgar, Cheltenham, 2016, p. 385.
93 F. Sauer and N. Schörnig, above note 41; Sarah E. Kreps, "Just Put It on Our Tab: War Financing and the Decline of Democracy", *War on the Rocks*, 28 May 2018, available at: https://warontherocks.com/2018/05/just-put-it-on-our-tab-21st-century-war-financing-and-the-decline-of-democracy/.
94 Denise Garcia, "Killer Robots: Toward the Loss of Humanity", *Ethics and International Affairs*, April 2015, available at: www.ethicsandinternationalaffairs.org/2015/killer-robots-toward-the-loss-of-humanity/.
95 DoD, above note 16.
96 E. B. Kania, *Battlefield Singularity*, above note 46.

https://doi.org/10.1017/S1816383120000466 Published online by Cambridge University Press

KRC survey data.[97] This conforms with earlier online polling conducted by the Open Roboethics Initiative[98] as well as Heather Roff via IPSOS.[99] Opposition in the United States, China and Russia is at 52%, 59% and 60% respectively.[100] In Europe, the numbers range from 60% in Finland up to 81% in Ireland.[101]

Survey data also suggest that the public's opposition is primarily fuelled not by legal concerns or worries about unwanted escalation or crisis instability but by the notion that delegating decisions over life and death on the battlefield crosses a bright red moral line.[102] So while there is certainly an interesting philosophical debate to be had about the cultural pervasiveness of human dignity as a concept and its relevance to the LAWS issue from utilitarian versus deontological viewpoints, the concern as presented in the preceding section quite clearly resonates with the general public. The notion that there is something fundamentally wrong with having humans killed by mindless machines is thus well suited to creating grassroots pressure on governments in order to muster more political will on the issue. This point is granted even by sceptics of the human dignity argument as a whole: "There could be some campaigning advantages. Saying that something is against human dignity evokes a strong visceral response."[103]

Fostering norm development in the CCW

LAWS keep steadily gathering media attention around the globe.[104] With mounting public pressure and increased scrutiny, there will be a strong incentive for CCW States Parties to produce tangible results for the 2021 Review Conference. The "aspects of the normative and operational framework" that are to be further developed over the course of 2021 could take a more concrete shape in three steps.

First, consensus seems achievable on shared language that adopts the by now widely accepted functionalist view of weapon autonomy as well as a common understanding that some form of positive obligation and affirmation of the principle of human control over weapons systems is required.[105] The CCW's guiding principle (b) already points this way in stating that "[h]uman

97 KRC, "Global Poll Shows 61% Oppose Killer Robots", 22 January 2019, available at: www.stopkillerrobots.org/2019/01/global-poll-61-oppose-killer-robots/.

98 Open Roboethics Institute, "The Ethics and Governance of Lethal Autonomous Weapons Systems: An International Public Opinion Poll", 9 November 2015, available at: www.openroboethics.org/wp-content/uploads/2015/11/ORi_LAWS2015.pdf.

99 Heather M. Roff, "What Do People Around the World Think about Killer Robots?", *Slate*, 8 February 2017, available at: https://slate.com/technology/2017/02/what-do-people-around-the-world-think-about-killer-robots.html.

100 KRC, above note 97.

101 KRC, "New European Poll Shows Public Favour Banning Killer Robots", 13 November 2019, available at: www.stopkillerrobots.org/2019/11/new-european-poll-shows-73-favour-banning-killer-robots/.

102 KRC, above note 97.

103 A. Sharkey, above note 85, p. 9.

104 R. Charli Carpenter, *"Lost" Causes: Agenda Vetting in Global Issue Networks and the Shaping of Human Security*, Cornell University Press, Ithaca, NY, 2014, pp. 88–121.

105 Stephen D. Goose and Mary Wareham, "The Growing International Movement Against Killer Robots", *Harvard International Review*, Vol. 37, No. 4, 2016, available at: www.jstor.org/stable/26445614.

https://doi.org/10.1017/S1816383120000466 Published online by Cambridge University Press

responsibility for decisions on the use of weapons systems must be retained since accountability cannot be transferred to machines".[106] The Forum for Supporting the 2020 GGE on LAWS conducted in April 2020 as a webcast by the German Federal Foreign Office, with 320 registered participants representing sixty-three CCW States Parties, underlined the importance of further conceptualizing the human element. Controllability of weapons is arguably a proto-norm already,[107] and a shared terminology – be it "meaningful human control" or some other formulation – could be found to stipulate in a general sense when humans and when machines are to be performing which function in the targeting cycle. The ICRC and the Stockholm International Peace Research Institute (SIPRI) recently presented a conceptual framework that can support this effort of operationalizing human control – that is, of clarifying the "who, what, when and how" of controlling weapons and limiting their autonomy.[108]

Second, since there is no one-size-fits-all standard of meaningful human control, the sharing of best practices and, more importantly, of case studies of specific weapons systems and operational scenarios could allow CCW States Parties to develop a deeper, shared conceptual grasp of the intricacies involved with implementing human control in design and use. The GGE is uniquely suited to facilitate these sorts of deep dives with analyses from multiple stakeholders and a sharing of legal, ethical and operational views. Smaller expert groups such as the International Panel on the Regulation of Autonomous Weapons (iPRAW) and the commission on the responsible use of technologies in the Franco-German Future Combat Air System are already beginning to organize their research toward that end.

Third, a differentiated implementation scheme could be developed that conceives of human control as being exercised in a context-dependent way – that is, contingent on the weapons system, its mission environment, "target profiles"[109] and additional factors such as mission duration.[110] This human control scheme could prescribe minimum standards for controllability by design, for example regarding the ergonomics of human–machine interfaces, and determine "levels of human supervisory control"[111] in use – that is, the tactics, techniques and procedures required to keep human control and responsibility intact during the system's operation.

106 CCW Meeting Final Report, above note 2, p. 10.

107 For the notion of codifying human control as a principle of IHL in general, see Elvira Rosert, *How to Regulate Autonomous Weapons*, PRIF Spotlight 6/2017, Peace Research Institute Frankfurt, 2017, available at: www.hsfk.de/fileadmin/HSFK/hsfk_publikationen/Spotlight0617.pdf.

108 V. Boulanin *et al.*, above note 26. See also Ilse Verdiesen, Filippo Santoni de Sio and Virginia Dignum, "Accountability and Control over Autonomous Weapon Systems: A Framework for Comprehensive Human Oversight", *Minds and Machines*, 2020, available at: https://link.springer.com/article/10.1007/s11023-020-09532-9.

109 Moyes, "Target Profiles", above note 26.

110 For this general approach as well as a list of variables to consider, see V. Boulanin *et al.*, above note 26, pp. 30–33.

111 F. Sauer *et al.*, above note 79, pp. 42–45.

https://doi.org/10.1017/S1816383120000466 Published online by Cambridge University Press

It currently seems unlikely that the CCW process, even if it were to complete these three steps, will end up yielding more than "soft law", such as a consensual political declaration or a catalogue of best practices. In fact, a complete breakdown of the CCW process in Geneva is also within the realm of possibility. But even if the CCW turns out not to be the venue from which a legally binding regulation for weapon autonomy emerges, it has already served as an information hub and norm incubator for the last six years – and will continue to do. Especially considering the effect of the COVID-19 crisis on meeting schedules around the globe, it is currently too early to tell if other fora – and if so, which ones – can and should pick up the ball on regulation where the CCW leaves it in 2021, in order to further develop and codify the human control norm as binding international law.

Conclusion

A multilateral regulation of autonomy in weapons systems – that is, codifying a legally binding obligation to retain meaningful human control over the use of force – is difficult yet imperative to achieve. Severe strategic as well as ethical mid- and long-term risks, such as unintended conflict escalation at machine speed and the violation of human dignity, outweigh any short-term military benefits. This analysis has illustrated how regulating weapon autonomy is feasible, presenting a three-step process to facilitate stepping back from the brink: step one, foster the emerging consensus on the notion that a positive obligation to retain human control over weapons systems is prudent and urgently required; step two, further develop the insight that there is no one-size-fits-all standard of meaningful human control; and step three, devise differentiated, context-dependent human control schemes for weapons systems. Given the current geopolitical landscape and the lack of political will to engage in arms control efforts, the taking of these steps will resemble a marathon, not a sprint. After all, the perceived military value of weapon autonomy is exceptionally high, and the issue itself is elusive, requiring an innovative, qualitative approach to arms control.

But history clearly suggests that great powers are not devoid of sensitivity to the accumulation of collective risks – otherwise arms control on nuclear, chemical and biological weapons would never have seen the light of day. The emerging technologies of the twenty-first century present humankind with the opportunity to demonstrate that it has learned from history *before* the risks have manifested themselves to their full extent. Humans do terrible things to each other in war, and there is no technological fix for that. But the international community can at least set rules to curb against uncontrolled escalation and the crossing of fundamental moral lines. If we fail to do so, we will not only lose

https://doi.org/10.1017/S1816383120000466 Published online by Cambridge University Press

the breathing room to ponder and deliberate responses,[112] an essential requirement of political conflict management, as the Cuban Missile Crisis strongly suggests;[113] we will also allow "the ultimate indignity" of war turning into "death by algorithm".[114]

112 Z. Davis, above note 35, p. 122.
113 Frank Sauer, *Atomic Anxiety: Deterrence, Taboo and the Non-Use of U.S. Nuclear Weapons*, Palgrave Macmillan, London, 2015, pp. 91–92.
114 Robert H. Latiff and Patrick J. McCloskey, "With Drone Warfare, America Approaches the Robo-Rubicon", *Wall Street Journal*, 14 March 2013, available at: https://tinyurl.com/y2t7odsh.

https://doi.org/10.1017/S1816383120000466 Published online by Cambridge University Press