Banning Autonomous Weapons: Legal, Ethical and Technical Challenges

Andrew Harrison[1], Hongyu He[2] and Serghei Mihailov[3]

[1]Universiteit van Amsterdam
[2]Vrije Universiteit Amsterdam
[3]Amazon Inc.

There is a reciprocal relationship between war and technology, with new technology changing warfare, and the exigencies of warfare driving new technological development (Leveringhaus, 2016). Even for technologies seemingly decoupled from war, a dual usage concern holds, as the same technology that guides automated cars to their destination can guide automated weapons to theirs (ibid). The use of Artificial Intelligence (AI) in military weaponry, referred to as Autonomous Weapons, or more emotively by Sparrow (2007) as 'Killer Robots', sits within a legally and ethically contested space. In 2013, a UN Special Rapporteur published a report on Lethal Autonomous Robots, calling for both national level moratoriums on their development, and the establishment of a high-level panel for the international community to start establishing a policy (Heyns, 2013). From 2014, the preexisting U.N. Convention on Certain Conventional Weapons (CCW) started discussing autonomous weapons (Scharre, 2018). By 2016, this led to the creation of the Group of Governmental Experts (GGE) sitting within the CCW (U.N.O.G., n.d.). The GGE met for the first time in November 2017 to discuss what had been termed Lethal Autonomous Weapon Systems (LAWS) (ibid). Prior to the GGEs' third meeting in March 2019, the UN Secretary General stated "machines with the power and discretion to take lives without human involvement are politically unacceptable, morally repugnant and should be prohibited by international law" (U.N., 2019). NGOs including Human Rights Watch, Amnesty International, and Article 36 are among the steering members of the Campaign to Stop Killer Robots, which is a "coalition of non-governmental organizations (NGOs) that is working to ban fully autonomous weapons and thereby retain meaningful human control over the use of force"

(Campaign to Stop Killer Robots, n.d.). In addition, prominent academics and public figures, along with private organisations, have signed the Lethal Autonomous Weapons pledge hosted by the Future of Life Institute, pledging that they "will neither participate in nor support the development, manufacture, trade, or use of lethal autonomous weapons" (Future of Life Institute, n.d. A). However, despite support from many national governments, particularly those in the global south, other governments have resisted, amongst them Australia, Israel, Russia, the UK, and the US (Gayle, 2019). Paradoxically, the UK Ministry of Defence states "the United Kingdom does not possess fully autonomous weapon systems and has no intention of developing them. We believe a preemptive ban is premature as there is still no international agreement on the characteristics of lethal autonomous weapons systems" (ibid.). The seeming logical fallacy of being able to neither possess nor develop something for which there are no actual defined characteristics, is political doublespeak, and when added to the inherent technical complexity of AI and advanced weaponry, is indicative of the confusion surrounding autonomous weapons. It is against this backdrop, that this essay highlights the pressing contemporary debates over banning autonomous weapons. It first examines the applicability of international humanitarian law (IHL) to, along with ethical question about, their usage. The subsequent technical aspect of the essay focuses on Automatic Target Recognition (ATR), for as Scharre states "the defining feature of autonomous weapons is how target selection and engagement decisions are made" (p.302, 2018). Finally, by tieing the technical examination back into the legal and ethical, and taking a historical perspective on the efficacy and feasibility of bans, a holistic prognosis for a possible ban on autonomous weapons is made.

Sharkey (2018) categorises the objections to autonomous weapons into three main types: (1) non-adherence with IHL; (2) deontological, based on human judgement and control, and including human dignity; (3) consequentialist, based on global stability, and increased likelihood of war. Further arguing that AI researchers may be more convinced by the first, lawyers and philosophers by the second, and politicians by the third. Similarly, Human Rights Watch (n.d.) frames the debate as being "questionable that fully autonomous weapons would be capable of meeting international humanitarian law standards, including the rules of distinction, proportionality, and military necessity, while they would threaten the fundamental right to life and principle of human dignity". According to Sharkey (2008), the International Humanitarian Laws most applicable to Autonomous weapons are, firstly the principle of discrimination; being able to make the distinction between on one hand combatants, and on the other hand, non-combatants or combatants that are surrendering, already captured, mentally ill or physically in-

jured. Secondly, the principle of proportionality; that loss of life and damage to property must be proportional to the direct military gain. Neither of these, making a distinction, nor calculating proportionality, can adequately be performed by autonomous weapons, and this needs to be addressed before their "inevitable proliferation" (p.89, ibid). As regarding military necessity, which is applicable to developing new weapons, often exemplified by the ban on explosive or combustible munitions under 400 grams (where a bullet already suffices) (Scharre, 2018), defensive weapons with capabilities for full autonomy have already been developed as they are required for flooding or overwhelming attacks; missiles, boats, or in the future drone swarms.

Within the ethical debates, Arkin (2010) argues for the ethics of autonomy on deontological and consequentialist grounds. Stating that autonomous weapons may be able to perform more ethically than humans, by strictly adhering to the laws of war, and by reducing non-combatant casualties and property damage. Countering this, is Sparrow's (2007) main premise that allocation of responsibility is required for jus in bello. Both for deontological, amongst others a Kantian respect for persons, and consequentialist, namely the lack of a prosecutable person leading to a propensity for war. Someone must be held accountable, however neither the programmer, commander, nor robot itself fits this mould, therefore it is unethical to deploy autonomous weapons. This is termed the accountability gap, and has become a key argument against autonomous weapons. Of the ten possible guiding principles agreed by the GGE in their 2018 meeting, principle two was based on this accountability gap, stating that "human responsibility for decisions on the use of weapons systems must be retained, since accountability cannot be transferred to machines" (Møller, 2019).

Delving further into the deontological, Sharkey (2018) examines and summarises the human dignity debate within autonomous weapons. Concluding that while human dignity may be grounds to support a ban on autonomous weapons, the core problems with its use are two-fold. First, is that what dignity means "varies between cultures, contexts, historical era, and philosophical position", and that its use may be more in the "campaigning advantages" from the "strong visceral response" it evokes (p.9, ibid). Secondly, establishing why autonomous weapons are different from other reducers of human dignity, such as war itself, other weapons, or human behaviour, in how they each affect dignity, is also fraught. Therefore, human dignity should not be relied on solely, nor too heavily, in arguments against autonomous weapons (ibid). Though not said explicitly, the same discernibility critique can be made to the right to life.

Efforts are being made to build a confluence model to provide support for a ban, combining a Prioritization of Deontology rule that deontological takes

primacy where applicable, with a default rule of applying consequentialist arguments when deontological are not applicable (Amoroso & Tamburrini, 2017). However, some research practitioners are also separating technological development from the ethics of actual deployment. The Director of the Tactical Technical Office (TTO) within DARPA's (Defence Advanced Research Projects Agency) states that their research is to provide the option, to "take that technical question off the table", and that the decisions about deployment and use are not theirs to make (p.83, Scharre, 2018). Which leads into the technical examination of autonomous weapons.

Automatic target recognition (ATR) is not just the most technically challenging aspect of autonomous weapons, it is also the most important aspect. Ekelhof (2018) believes that within "autonomous weapons, we should focus our attention first and foremost on what should be considered targeting" (p.1, 2018). Scharre also notes that it is "important to separate the effects of robotics and automation in general from autonomous targeting in particular" (p.302, 2018). ATR offers enhanced automated data analysis to analysts, which then directly affects their decision-making. A reliable ATR system can dramatically improve target lethality and non-target survivability, by reducing human workload and offering automated cues for action in a combat situation (Blacknell & Vignaud, 2013; Ratches, 2011).

At the core of ATR is image processing and machine learning. The following analysis describes Air-to-Ground (ATG), which is air-based ATR of ground-based targets. The first step is gathering image data via unmanned aerial vehicles (UAVs). Novak et al. describes, "its [UAV] mission goals include an operating range of 3,000 nautical miles and the ability to loiter over the target area for 24 hours at altitudes of 65,000 feet. In addition to electro-optical and infrared sensors, this UAV will carry a SAR sensor that is projected to collect in one day enough data sampled at a resolution of 1.0 m $\times$ 1.0 m to cover 140,000 km2 (roughly the size of North Korea)" (p.187, 1997). The ATR preprocesses the images, roughly labelling the possible targets according to abnormal shadows and specular returns. A clustering algorithm is applied to distinguish the potential targets from the background, with the ATR then extracting information about the targets, paving the way for the next step, classification.

Classification is the key step in ATR, determining the accuracy and efficiency of the entire process. Two widely used classification algorithms are template matching and feature-based classification. For higher accuracy, ATRs using template matching classification require 360-degree image data to already be stored, a major burden on databases that decreases the time performance of the ATR system (Blacknell & Vignaud, 2013). In feature-based classification, establishing different classes of targets that have com-

pletely separate feature vectors is challenging. There is always partial overlap in target features, meaning that these ATR systems can be confused and misclassify in certain circumstances (ibid). After the classification step, the image data is cued and ready for analysis by the image analysts. These ATR systems can reduce human labour to around three people (two image analysts and a supervisor, from what was previously seven image analysts and one supervisor) and is able to achieve an almost real time efficiency (around five minutes, when previously it took thirty minutes) (Novak et al., 1997).

The question then arises, why should an ATR system need human intervention at all? There are several challenges that prevent ATR systems for ground targets from being fully-autonomous, rather than just providing human decision support. The number and type of ground targets is enormous, and there is a myriad of different environments that surround the targets. The performance of ATR systems drops dramatically when the complexity of targets and environments increase (Blacknell & Vignaud, 2013). State-of-the-art ATR algorithms (El-Darymli et al., 2016; Goodwin et al., 2018; Chen et al., 2016) show an accuracy of 97% in standard operating conditions, but drop severely in accuracy (to 70% and below) in the presence of noise in the image, where there is occlusion of the object to be recognized, or where there is a lack of views of the object from multiple perspectives.

State-of-the-art image classification systems have been shown to perform on par with trained humans (Russakovsky et al., 2015), with 6.8% and 5.1% error rates on the ImageNet challenge, respectively. But the former does not compare trained human classifiers in cases of noisy data, which is often the case for battlefield data: "by increasing the noise width from 0.0 (no noise) to 0.1, VGG-16's performance drops from an accuracy of 89.91% to 44.02%; GoogLeNet's drops from 81.70% to 34.02% and AlexNet's from 70.00% to 19.29%. Human observers, on the other hand, only drop from 80.50% to 75.13%." (p.8, Geirhos et al., 2017). Another challenge is that there is always a trade-off in ATR performance. Higher detection rates also mean higher false alarm rates, which is dangerous in battlefields in the sense of firing at the wrong target and also disclosing the firing platform's location (Ratches, 2011). There is also a trade-off needed when seeking high-speed and high-accuracy at the same time (Blacknell & Vignaud, 2013). In addition, ground targets are always moving creating slight nuances that have a significant impact on present ATR systems (ibid). In conclusion, ATR technology can improve weapon systems' efficiency but is not currently able to support fully-autonomous weapons.

Based on the legal, ethical, and technical aspects of autonomous weapons described above, a discussion on the feasibility of banning autonomous weapons is warranted. Scharre (2018) believes that achieving a ban depends on three

things; the horribleness of the weapon, its military utility, and the number of parties required to cooperate in achieving the ban. Regarding feasibility, there are mixed results with banning weapons, and it can revolve around definition. Scharre (2018) records that asphyxiating gases (from projectiles) were banned in the 1899 Hague Declaration, but failed in WW1, as the German's argued that they were using 'canisters' rather than projectiles. Gases were then not used in WW2, despite, or perhaps given the horribleness and utility of the mechanised slaughter that signified that war. But more contemporary dictators, first in Iraq and then Syria have used gases again. Other recent successes have been the 1997 Mine Ban Treaty (Ottawa Treaty), and the 2008 Convention on Cluster Munitions. However, in both, exceptions were made by defining out of scope of the bans, weapons that countries had recently developed (smaller cluster munitions), or still valued (anti-vehicle mines with personnel anti-tampering fitted) (ibid). The definition issue is important, for example, Scharre (2018) takes umbrage with the Future of Life Institute's (n.d. B) Open Letter on Autonomous Weapons, currently signed by over 4,500 AI/Robotics researchers. Specifically, with banning "offensive autonomous weapons beyond meaningful human control" (ibid), stating that "every single one of those words is a morass of ambiguity" (p. 353, Scharre, 2018), and that if states could even agree on offensive and defensive definitions, then offensive weapons would have already been banned. As described above, without explicit encompassing definitions, countries can flout the bans. But even with strict definitions, countries put flaws in the bans.

Another feasibility issue, is how to ban something that, as the UK Government states, has no internationally agreed characteristics. A possible solution is the preemptive ban. Though the issue with preemptive bans, is that the eventual effects of the technology are hard to discern (Scharre, 2018). Scharre (2018) details attempts to ban submarines, first at the 1899 Hague Convention, and then 1921-1922 Washington Naval Conference, both failed to be ratified. However, strictly regulating their use was successful at the 1907 Hague Declaration, followed by the 1930 London Naval Treaty, and the 1936 London Protocol. But, adherence to these restrictions collapsed quickly during WW2, providing lesson one, which is the primacy of military utility during times of war. The second lesson is perhaps more revealing. Submarines are in wide use today, and arguably provide international stability against nuclear war, as roving nuclear armed submarines remove the decisive advantage of a nuclear first strike that totally obliterates a nation's land-based nuclear retaliation. Because "autonomous weapons [now] raise important issues for stability" (p.351, Scharre, 2018), what then might be their impact. One outcome is that autonomous weapons may reduce the

'body bag count', thus lowering the cost of going to war, and therefore making it more likely (Sharkey, 2018). As "technology will evolve in unforeseen ways. Successful preemptive bans [will] focus on the intent behind a technology, rather than specific restrictions" (p.343, Scharre, 2018). This fits with the GGE 2018 statement above, on human responsibility needed for decision on weapons usage. Perhaps unsurprising, given that Scharre receives special thanks in the United Nations Institute for Disarmament Research primer for CCW delegates (UNDIR, 2018).

Turning to the capacities that an autonomous system must have, Scharre (2018) differentiates between the three roles that a human agent plays: essential operator; fail-safe, and moral agent. Automating the essential operator component is the easiest, with clearly defined benefits. However, the fail-safe and moral components are far more challenging, and beyond the current capacity of AI technology. The human agent thus seems critical, with the former U.S. Deputy Secretary of Defence describing the need for "the human always in front... that's the ultimate circuit breaker" (p. 228, Scharre, 2018). As for performing moral agency, it is suggested "that 'death by algorithm' crosses a moral line" (p.10, Sharkey, 2018) itself, therefore appearing irredeemably immoral.

That horribleness may override the military utility is not sufficient, as the fractious parties required to implement a ban must still agree. However, the summation from Acheson (2019) of the March 2019 GGE talks was thus, "another round of UN talks on autonomous weapon systems ended ... without significant movement in any particular direction. Six years into this political process, states are continuing to tread water". Scharre (2018), delineates four possible outcomes that may be reached: 1) banning autonomous weapons, 2) banning anti-personnel autonomous weapons, 3) creating 'rules of the road' for using autonomous weapons, 4) creating a new general principle on human judgement's role in war. The fourth ties back into pre-emptive bans on the intent behind a technology, the intent behind AI is to replicate intelligence and thus in autonomous weapons have autonomous decision-making. Thus, codifying the need for human judgement in war through a new principle, appears to be the most feasible means to effectively implement a ban on autonomous weapons.

However for a ban, social and political will is still required. A January 2019 poll across 26 countries, with 500 - 1,000 respondents per country, showed an increase since 2017 (from 56% to 61%) in the percentage of the general public that opposes the development of Lethal Autonomous Weapons (Amnesty International, 2019). While democratic governments are susceptible to the will of the people, they are also responsible for the economic and physical security of their nation. As Scharre states, "the main rationale for

building fully autonomous weapons seems to be the assumption that others might do so" (p.330, 2018). Take the EU as a microcosm to examine the political to-and-fro over banning autonomous weapons. In September 2018 the EU Parliament passed a resolution calling for "international negotiations on a legally binding instrument prohibiting lethal autonomous weapons systems" (Campaign to Stop Killer Robots, 2019). Then, at the European Defence Agency's annual conference in November 2018, the EU Foreign Affairs chief stated "almost 50 percent of global private investment in artificial intelligence startups is happening in China. We Europeans cannot afford to waste time and to be less innovative than other world powers. It is a matter of economic growth, and it is a matter of security" (Banks, 2018). Followed in February 2019 by a provisional agreement (requiring EU Council and Parliament approval), that the EU Defence Fund's (EDF) budget for 2021-2027 could not provide funding for autonomous weapons (Campaign to Stop Killer Robots, 2019). However, and it is a large however, as seen before with landmines and cluster munitions. The EDF agreement has qualifiers, allowing funding for anti-material (as opposed to personnel) and defensive autonomous weapons (ibid). Ultimately, the economic and security implications of AI create the conditions for a race to the ethical bottom. In concluding, that autonomous weapons may one day be able to confirm with IHL, is currently debated, but if possible, it would then leave only the deontological and consequentialist arguments for their ban (Sharkey, 2018). Humanity is thus left to weigh the horribleness of 'computer says kill', against military utility. Are Nagasaki and Hiroshima the exceptions that prove the rule, or the rule that we will always make horrible exceptions.

References

Acheson, R. (2019, April 1). Preventing a march toward dystopia. Retrieved April 30, 2019.

Amnesty International. (2019, January 22). Killer robots: New global poll shows growing public opposition to autonomous weapons. Retrieved April 27, 2019.

Amoroso, D., & Tamburrini, G. (2017). The ethical and legal case against autonomy in weapons systems. Global Jurist, 18(1).

Arkin, R. C. (2010). The case for ethical autonomy in unmanned systems. Journal of Military Ethics, 9(4), 332-341.

Banks, M. (2018, November 30). EU members seek common ground on autonomous weapons. Retrieved April 27, 2019.

Blacknell, D., & Vignaud, L. (2013). Atr of ground targets: Fundamentals and key challenges. RADAR Automatic Target Recognition (ATR) and Non-Cooperative Target Recognition (NCTR), 1-1, 1-36.

Campaign to Stop Killer Robots. (n.d.). About. Retrieved April 27, 2019.

Campaign to Stop Killer Robots. (2019, February). No killer robots for European Defence Fund. Retrieved April 30, 2019.

Chen, S., Wang, H., Xu, F., & Jin, Y. Q. (2016). Target classification using the deep convolutional networks for SAR images. IEEE Transactions on Geoscience and Remote Sensing, 54(8), 4806-4817.

Ekelhof, M. A. (2018). Lifting the Fog of Targeting:"Autonomous Weapons" and Human Control through the Lens of Military Targeting. Naval War College Review, 71(3), 6.

El-Darymli, K., Gill, E.W., Mcguire, P., Power, D. & Moloney, C. (2016) "Automatic target recognition in synthetic aperture radar imagery: A state-of-the-art review." IEEE Access 4 (2016): 6014-6058.

Gayle, D. (2019, March 29). UK, US and Russia among those opposing killer robot ban. Retrieved April 27, 2019.

Geirhos, R., Janssen, D. H., Schütt, H. H., Rauber, J., Bethge, M., & Wichmann, F. A. (2017). Comparing deep neural networks against humans: object recognition when the signal gets weaker. arXiv preprint arXiv:1706.06969.

Goodwin, J. A., Brown, O. M., Killian, T. W., & Son, S. H. (2018). Learning Robust Representations for Automatic Target Recognition. arXiv preprint arXiv:1811.10714.

Heyns, C. (2013). Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions, A/HRC/23/47, United Nations General Assembly, 9 April.

Human Rights Watch. (n.d.). Killer Robots. Retrieved April 29, 2019.

Future of Life Institute. (n.d. A). Lethal Autonomous Weapons Pledge. Retrieved April 27, 2019.

Future of Life Institute. (n.d. B). Open Letter on Autonomous Weapons. (n.d.). Retrieved April 30, 2019.

Leveringhaus A. (2016) Ethics and the Autonomous Weapons Debate. In: Ethics and Autonomous Weapons. Palgrave Pivot, London.

Møller, M. (2019, March 25). Secretary-General's message to Meeting of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems. Retrieved April 30, 2019.

Novak, L. M., Owirka, G. J., Brower, W. S., & Weaver, A. L. (1997). The automatic target-recognition system in SAIP. Lincoln Laboratory Journal, 10(2), 187-202.

Ratches, J. A. (2011). Review of current aided/automatic target acquisition technology for military target acquisition tasks. Optical Engineering, 50(7), 072001/1-072001/8.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. International journal of computer vision, 115(3), 211-252.

Scharre, P. (2018). Army of none: Autonomous weapons and the future of war. WW Norton & Company.

Sharkey, A. (2018). Autonomous weapons systems, killer robots and human dignity. Ethics and Information Technology, 1-13.

Sharkey, N. (2008). Grounds for discrimination: Autonomous robot. RUSI Defence Systems, 11, 86–89.

Sparrow, R. (2007). Killer robots. Journal of applied philosophy, 24(1), 62-77.

U.N. (2019). Autonomous weapons that kill must be banned, insists UN chief | UN News. (2019, March 25). Retrieved April 26, 2019.

UNDIR. (2018). The Weaponization of Increasingly Autonomous Technologies: Artificial Intelligence (a primer for CCW delegates). Retrieved from April 26, 2019.

U.N.O.G. (n.d.). 2019 Group of Governmental Experts on Lethal Autonomous Weapons Systems (LAWS). Retrieved April 27, 2019.