ARTICLE

MIT Science Policy Review

Lethal autonomous weapons systems & artificial intelligence: Trends, challenges, and policies

Shayne Longpre[1,*], Marcus Storm[2], and Rishi Shah[3]

Edited by Kevin McDermott

HIGHLIGHTS

- Autonomous weapons are rapidly proliferating: in accessibility, their degrees of autonomy, the range of international developers, and their tactical roles in intelligence, reconnaissance, and lethal strikes.
- Autonomous systems remain highly prone to error, demonstrating poor robustness, interpretability, and adversarial vulnerability.
- Major military powers abstain from treaty proposals, while other nations and humanitarian organizations urgently demand regulation.
- International and U.S. policy remains ambiguous and lacks realistic accountability and enforcement mechanisms.

We see and expect increased global proliferation of lethal autonomous weapons. Global coordination is needed to control and regulate these weapons.

This paper surveys the key technical, humanitarian, and political challenges faced by the global community in the proliferation of autonomy in lethal weapons systems. The discussion herein covers weapons systems with varying types of autonomy, and in particular lethal autonomous weapons systems (LAWS): "weapon systems that, once activated, can select and engage targets without further intervention by a human operator" [1], which include armed drones, vehicles, submersibles, sentry turrets, missile systems, and other *kinetic* applications of artificial intelligence (AI). This report aims to summarize the key developments from the public domain, without clandestine information. Please contact the authors with corrections and additions.

We first discuss the state of LAWS, including general trends, a timeline of LAWS, degrees of autonomy, the dual-use nature of some of the technologies used in LAWS, international political stances, and challenges presented by accessibility, interpretability, accountability, and enforcement. Next, we discuss recent literature in autonomous capabilities and intelligence for LAWS, followed by a discussion of autonomous weaponry policy from domestic and international perspectives.

Terminology & Acronyms	
AI	Artificial Intelligence
LAWS	Lethal Autonomous Weapons Systems
UAS/UAV	Unmanned Aerial System / Vehicle
RPA	Remote Piloted Aircraft
USV	Unmanned Surface Vehicle
ISR	Intelligence, Surveillance, and Reconnaissance

Global trends

In recent years, technological advances in hardware, from electro-optical, infrared, and sonar systems to synthetic aperture radar, and in AI/robotics, from better 3D and visual perception to motion forecasting and planning, have enabled the rapid development of autonomous systems [2]. Technological advances drive lower costs, greater accessibility, less human error, and accelerated reaction times; this expands the opportunities for uses in intelligence, surveillance, and reconnaissance (ISR), navigation, detection, and so on. AI-enabled devices provide unmanned vehicles with greater speed and the ability to operate in environments without the need for data links, e.g. underwater or near adversary jamming devices, increasing opportunities to outmaneuver enemy systems [3]. Recent trends see a wider set of nations involved in active LAWS development, including increasingly offensive applications, positioning for urban conflict rather than battlefields, and swarming capabilities. Additionally, the capabilities of unmanned aerial vehicles (UAVs) are expanding in their duration of deployment, geographical areas of operation, their range of identifiable objects, and ability to coordinate among themselves [2].

The international debate has intensified correspondingly. Nation states through the United Nations, together with individual citizens through international advocacy groups, apply increasing pressure to impose legally binding

[1]Massachusetts Institute of Technology, Cambridge, MA
[2]Imperial College London, London, UK
[3]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA
*Email: slongpre@mit.edu

The authors declare no conflict of interest.

Selected Aerial Military Systems with Autonomous Capabilities

Type	Name	Developer	Country	Usage	Autonomy	Year
Loitering Munitions	Drone 40	DefendTex	Australia	Quadcopter + grenade	Nav + Target	2016
	Mini Harpy	Israel Aerospace Industries	Israel	Mini-UAS + munition	Nav + Target + Fire	2019
	KUB-BLA	Kalashnikov	Russia	Loitering munition	Nav + Target + Swarm	2019
	Kargu	STM	Turkey	Loitering munition	Nav + Target + Fire	2020
Unmanned Aerial Vehicle	Bayraktar TB2	Bayraktar	Turkey	Unmanned aerial combat vehicle	Nav	2014
	MQ-9 Reaper	General Atomics	U.S.	ISR	Nav + Identify	2020
	Unnamed SRR drone	Skydio	U.S.	ISR	Nav + Identify	2022
Aircraft	Ghost Bat	Boeing	Australia	Wingman UAS controlled by manned parent	Nav + Target + Fire	2020

TABLE I: Selected Aerial Military Systems with Autonomous Capabilities. Sourced from [2] and publicly available knowledge.

international treaties regulating such weapons; yet the largest military actors, including the United States, have repeatedly refrained from any such commitments. Consequently, the technological advancement and military adoption of a wide variety of LAWS with increasingly autonomous functions continues forward.

Timeline & examples of LAWS: LAWS have been evolving for decades. In prior decades, Semi-Automatic Ground Environment (SAGE) air defense systems searched for hostile jets, and warships employed close-in weapon systems (CIWS) to automatically detect, track, and eliminate incoming missiles [4, 5]. Perhaps the most famous example of human intervention in automated warfare was the 1983 Soviet nuclear false alarm incident, triggered by automatic target detection. Stanislov Petrov, a Soviet Air Defense lieutenant colonel, chose not to invoke the Soviet policy of compulsory nuclear counter-attack when early warning satellites incorrectly identified high-altitude clouds as intercontinental ballistic missiles traveling from the United States.

The 1980s saw the manufacturing and development of third-generation Anti-Tank Guided Weapons (ATGWs), which were designed to be fired upwards into the air and would acquire targets independently using infrared. The European PARS 3 LR [6] and Israeli Spike [7] ATGWs are examples of this type of homing missile. The modern Javelin ATGW, sent by the United States and other nations to the Ukrainian army in 2022, incorporates a control system called an electronic safe arm and fire (ESAF) which directs the missile towards the target after launch [8].

The U.S. has been a prominent innovator in the field of autonomous weapons, pioneering a target tracking and acquisition unit named Phalanx CWIS first produced in the 1970s [9]. In the early 2000s the U.S. Patriot Missile computers misidentified friendly jets on two separate occasions, which led to friendly fire and death [10]. Flawed procedures had not properly accounted for automation error. In the last decade, South Korea and Israel have built sentry guns capable of recognizing and firing on humans with complete autonomy [10].

In recent years, Russia and Israel have also developed unmanned surface vehicles (USVs) with autonomous navigation and targeting capabilities, and China has developed an autonomous helicopter [5, 11]. The advent of lethal UASs and loitering munitions are perhaps the most dangerous LAWS, now widely developed and relatively cheap and accessible [11]. Previously used for reconnaissance, these aerial systems are designed to autonomously patrol regions, search for enemy radar, aircraft or people, and intercept them, often with a built-in warhead. Numerous examples are shown in Table I.

On 8 March 2021, the Panel of Experts on Libya submitted UN Letter [S/2021/229] [12] to the United Nations Security Council. According to the report, on 27 March 2020, the forces of Khalifa Haftar were attacked by at least one Kargu-2 LAWS, documenting the first likely fully autonomous fire-and-forget usage of a lethal autonomous weapon. There may have been similar attacks yet unknown to the public due to the difficulty in confirming whether a weapon such as this was truly acting autonomously or not.

Russia's invasion of Ukraine in 2022 saw the widespread use of the TB2 (Table I) and Javelin ATGWs. It is not clear whether TB2 autonomy in takeoff and cruise functionalities have contributed to the war, but the Javelin missile's "fire-and-forget" capabilities have enabled small counterattacking forces to rapidly strike and retreat at distance [13].

Dual-use technology: Ref. [14] describe the "dual-use dilemma" of artificial intelligence: that the same technology offers both critical civilian and military applications. The same visual perception, human identification, and tracking tools which self-driving cars use to steer clear of pedestrians are easily re-tasked with finding and detonating on military targets.

Risks & Challenges	
Dual-Use Technology	Technologies for object identification, tracking, and navigation, which are critical for civilian applications, are often adapted for lethal military applications.
Accessibility	The hardware and software components for LAWS are increasingly affordable and accessible.
Accountability	Legal ambiguity and technological limitations create an accountability gap for the actions of autonomous agents.
Attribution & Compliance	Reduced traceability and interpretability for LAWS complicates attribution, regulation, and enforcement of laws governing conflict.
Ethics	The automation of violence, especially in targeting humans is ethically dubious.
Political Fragmentation	The United States and other prominent political players refuse to engage in negotiating international treaties regulating LAWS.
Interpretability	State-of-the-art machine learned systems offer little justifications or diagnostic tools for their decisions.
Generalization & Robustness	Machine perception, tracking, and navigation adapt poorly to unseen environments or circumstances.
Decision Making	The time between pre-programmed decision criteria and the autonomous attack escalates the risks of unintended consequences.
Adversarial Vulnerability	Machine learned systems are extremely vulnerable to intentional perturbations in physical environments.
Facial Recognition	The use of facial recognition, gait recognition, or phone sensing technologies are increasingly considered for automated identification and targeting in LAWS, as well as automated surveillance.

TABLE II: Risks and Challenges to the development and regulation of LAWS and semi-autonomous ISR systems.

Autonomous systems extend many positive benefits, from clearing land mines, supplying contested territories, identifying and safeguarding non-combatants, and limiting collateral damage. These applications rarely require automated targeting or firing. DART, the Dynamic Analysis and Replanning Tool, which used AI to optimize logistics and scheduling during Operation Desert Storm, is reported to have offset the expense of all DARPA funding for AI research in the prior 30 years [15]. Coupled with the ever-increasing accessibility of open source AI tools and technologies, disentangling harmful from beneficial applications may be more challenging than for nuclear, chemical, or biological weapons.

Shades of autonomy: The simplest form of autonomy is to enact decisions in an environment without human instruction – such as a land mine automatically triggering from contact. Ref. [16] discusses the dimensions of machine autonomy, including the human-machine command-control relationship, the sophistication of decision-making, and the autonomous function. For command-control relationships, many autonomous functions are largely assistive to humans – alerting of detected signals, identifying objects, suggesting schedules or routes, or aiding in targeting. Increasingly, though, this relationship is inverted, where the human merely assists the machine, with veto or override capacity [17]. Prior work reports human supervision over machines is rife with unsolved challenges [18]. In fully autonomous systems, the machine will conduct its own actions based on pre-programmed human instructions, or objective(s) with a set of constraints rather than a set of explicit instructions (i.e., "specification gaming" as described in Ref. [19]).

In terms of the autonomous function, many UAS and USVs incorporate some *autonomy of navigation*: to travel a route or arrive at a destination without human piloting. Navigation usually requires the ability to identify environmental conditions (obstacles, humans, other aircraft or vehicles) to navigate around. For *autonomy of identification* the machine detects and decides the identity or composition of its environment – required to some extent in navigation to avoid obstacles. *Autonomy of target selection ("targeting")* and *autonomy of firing*, are the most dangerous, allowing machines to choose their own targets and then initiating fire of their own accord. Only systems with these autonomous target selection and attack capabilities (critical functions) are considered LAWS.

Accessibility: Costs of lethal autonomous weaponry have been driven down by technological advances. Modern loitering munitions are also reusable if they have not detonated.

Each modern FGM-148 Javelin ATGW costs the U.S. Department of Defense $175k [20], and is large and difficult to assemble quickly and stealthily. In contrast, open-source image recognition and navigation software attached to cheap drones and homemade explosives suddenly make similar weapons – or fleets of these weapons – more accessible to actors with fewer resources.

Interpretability: Among modern machine learned systems, there exists an implicit trade-off between accuracy and the interpretability of a model's decision. Large artificial neural networks offer the best performance for most sophisticated tasks, including visual perception, forecasting, and motion planning, but most architectures currently provide limited human-understandable explanation for the model's output – effectively becoming a "black-box". Additionally, neural network models are often poorly calibrated, particularly in out-of-distribution environments, meaning their generated confidence scores can occasionally over- or under-exaggerate

their chance of error. In certain medical applications, automated systems are required to give some interpretable explanation of their decisions.

As an alternative, the United States requires a human-in-the-loop for any *lethal* systems [1]. However, it is unclear whether the speed and information available to these human agents is always sufficient to rectify automated mistakes. The documented phenomenon of "automation bias" shows humans begin to trust automated system excessively after extended use, biasing their judgement [21].

Accountability: The NGO Human Rights Watch affirms "an accountability gap" where "neither criminal law nor civil law guarantees adequate accountability" for actors involved in the chain of autonomous system design or command [22]. Ultimately, experts agree fully autonomous systems cannot feasibly inherit the liability from their designers, despite the replacement of some human decision-making. In the event of catastrophic errors in machine judgement, there is uncertainty whether engineers, product designers, users, or leadership teams are to be held responsible. To breach the Geneva Convention's Law of Armed Conflict requires some evidence that unlawful acts due to AI were *foreseeable* [5]. The lack of model *interpretability* may complicate this verdict, given inadequate explanations for the cause of events.

Attribution & enforcement: Related to the accountability gap, attacks carried out by LAWS complicates attribution and enforcement. Duplicitous actors may use autonomous weapons to reduce the traceability of their attacks, or blame autonomous system errors to disguise their intentions. Responses to seemingly unprovoked attacks will have to contend with the possibility of misrepresentation or misdirection, more easily concealed by autonomous systems. This poses a problem for deterrence. System hacks or denial of service attacks that cannot always be successfully traced serve as a cyberwarfare analog. Others have warned that without traceability, and particularly because LAWS are cheap and replaceable, robot warfare may engender more rapid escalation in future confrontations [23, 24].

Additionally, a broader problem for compliance is that without data access, it is difficult to prove if a weapon was operating autonomously. There are few extrinsic factors to separate an autonomously-acting weapon from a human-operated one.

Political stances for the regulation of LAWS: Over 70 nations have called for a fundamental ban on fully-autonomous weapons, with regulation on autonomy in weapons to ensure they comply with legal and moral standards. These nations include Argentina, Austria, Brazil Egypt, New Zealand, Norway, Pakistan, and Switzerland, among other nations (see Ref. [25] for the full list of nations). China has called for a treaty, while also investing heavily in LAW development.

Prominent humanitarian figures, NGOs, and advocacy groups also demand regulation for autonomous weapons, especially with regard to targeting and firing, including the United Nations Secretary-General António Guterres, Amnesty International, Human Rights Watch, the Campaign to Stop Killer Robots, itself a coalition of hundreds of organizations, and most recently the International Committee of the Red Cross (ICRC). Secretary-General António Guterres said "machines with the power and discretion to take lives without human involvement are politically unacceptable, morally repugnant" [26]. In lieu of full autonomy, the UN and other organizations have advocated for "meaningful human control" in the Convention on Certain Conventional Weapons (CCW) [27]. Similarly, the ICRC "recommends that states adopt new, *legally binding rules* to regulate autonomous weapon systems to ensure that sufficient human control and judgement is retained in the use of force. It is the ICRC's view that this will require prohibiting certain types of autonomous weapon systems and strictly regulating all others" [28]. Experts mention an eventual objective is to establish a legally binding treaty, or in the absence of full agreement, to stigmatize these weapons, thereby creating an international norm against their use.

Political stances against the regulation of LAWS: Notably absent among the nations calling for regulation of LAWS are eight powerful and globally engaged militaries: Australia, India, Israel, Russia, South Korea, Turkey, the United Kingdom, and the United States [29]. These nations actively invest in a growing arms race for technological superiority [25].

The U.S. military currently requires human-in-the-loop for any *lethal* autonomous weaponry [1]; however, active development areas seem to defy this stipulation: for example, there is funding available for swarming technology, such as a 2017 Pentagon research proposal requesting a "Cluster UAS Smart Munition for Missile Deployment" [30]. Ref. [10] argues that a large swarm of UAS designed to rapidly find and destroy targets could not feasibly be lethal with meaningful human control.

Military experts also argue the futility of a treaty or ban given the benefits of speed and protection conveyed by LAWS [31]. The Final Report from the National Security Commission on Artificial Intelligence, commissioned by Congress, even argues a ban would complicate enforcement for weapons already in the U.S. arsenal, and therefore recommend against it [32]. However, members of this commission have noted conflicts of interest [33].

Autonomous capabilities and intelligence

System design - hardware: LAWS are increasingly outfitted with more capable sensor hardware, which improve their autonomous perception functions. Aside from an array of optical cameras, USVs are frequently paired with 3D LiDAR, and UASs with 2D laser range finders. Newer candidates for inclusion are depth, light-field, and event-based cameras, as well as magnetic, olfaction, and thermal sensors [34]. The Kargu-2, a small UAS, is just 70x70x40cm and weighs 7kg. They contain carbon-fibre blades, a small computer chip, a portable controller, and house swappable explosives inside their main frame [35].

System design - algorithms for autonomy: Autonomous navigation remains a challenging and active area of development, propelled by the commercial prospect of self-driving cars. Typically, the objective is divided into four underlying tasks for artificial systems [36]–[38], each critical elements to navigation, targeting, tracking and other autonomous functions:

1) **Perception** Also referred to as "detection", this step predicts the presence of surrounding physical objects and their 3D bounding boxes, from the visual and range sensory signals.
2) **Object tracking** This step uses object detection temporally, over past frames. Linking together detected objects over time enables modeling of the trajectory, velocity, and, if applicable, the supposed intent of other mobile physical objects.
3) **Motion forecasting** This predicts the future positions for the set of objects tracked and resolved from previous frames.
4) **Motion planning** Given a representation of the scene, and the likely distribution of object motions, the vehicle plans its own motion in accordance with short- and long-term objectives: safely avoiding collisions, and reaching its destination.

Advances in hardware and machine learning have led to end-to-end neural network models, rather than cascade of specialist models, gaining greater performance at the expense of modularity, interpretability, and a human engineer's ability to impose intermediate interventions.

Perceptual generalization & robustness: There remain a few technological obstacles to more reliable or fully autonomous systems. The first is the lack of robustness in perception systems, which regularly fail to detect a real object, predict something which is not there, or incorrectly identify an object's type. Perception will often fail when confronted with environments dissimilar from what they were trained on, known as poor generalization [39, 40]. If the lighting conditions, distribution and appearance of surrounding objects, e.g., plants, buildings, people, or the behaviour of these objects are different from training, it may regularly lead to catastrophic errors. The presence of snow, fog, mirrors, birds, unfamiliar urban planning, or other new features of their environment are often sufficient to trigger errors [40]. Accurate perception, localization and scene understanding are critical to robust, collision-free motion planning in diverse conditions. Consequently, engineers prefer to train autonomous systems in the particular environments they plan to operate in, though this may not always be possible for real and unanticipated conflict zones, nor a guaranteed solution, as environments change during conflict, for instance, human behavior changes, erected barriers, and destroyed buildings. As a result, the data autonomous systems are trained on are almost never sufficiently complete, representative, and of high enough quality.

Further, operationalizing these systems outside of test environments come with many unforeseen socio-political risks aside from the technological ones described above [41]. And identification or targeting of humans is itself a poorly defined task [42].

Vulnerability to adversarial examples: In addition to natural errors, artificial neural networks are known to be susceptible to *intentionally* "adversarial" examples – minor input perturbations which may be imperceptible to a human [43, 44].

Prior literature has extensively demonstrated the dangers of visual and non-visual adversarial examples [45]–[47], and recently this has been demonstrated for *physical* adversarial examples [48]–[50]. The authors in ref. [48] successfully construct perturbations on physical objects that fool image classifiers under various real-world conditions. These physical perturbations can effectively fool state-of-the-art perception systems from as close as 30 feet, posing reliability challenges to autonomous systems in dense, human-populated environments.

Where visual perception vulnerabilities exist, 3D LiDAR sensors are usually seen as a defensive solution. However, [51] also demonstrate a physically realizable method to fool LiDAR detectors, generating a 3D object that when mounted on a vehicle renders them invisible to LiDAR detection systems with high accuracy.

Up to here we have discussed the robustness of visual or LiDAR-based perception. The authors in Ref. [52] successfully demonstrate attacks against the *end-to-end* autonomous systems, resulting in final *physical consequences* – in this case lane violations or crashes for USVs. These attacks are physically realizable, simple to implement, and appear inconspicuous to humans.

Altogether, autonomous systems face numerous unsolved challenges in the robustness, generalization, and vulnerability to attack. Though much literature is devoted to defenses, the unbounded nature of possible attacks means there has been no panacea. In many cases, only human control – slowing down the decision loops – can reliably diminish the potential for catastrophic error, e.g. target mis-identification or unrecognized presence of civilians.

Decision making: Aside from machine perception, LAWS are forced to make nuanced judgements on proportionality and distinction. Given their limitations, commanders need to pre-program these decisions in advance of deployment. The greater distance between the codified decisions and parameters and the actual attack pose significant hazards for unintended consequences.

Facial recognition: Manufacturers of military drones now offer integrated facial recognition software for automated target identification [35]. State-of-the-art facial recognition systems are built with the same basic ingredients as perception/detection systems described above and are therefore equally susceptible to their risks and challenges: lack of robustness, poor generalization, and adversarial

vulnerabilities.

The task of precisely identifying human facial features must contend with uncontrolled illumination, occlusion, pose variations, variability in facial expressions, makeup, facial hair, and clothing [53, 54]. Under uncontrolled settings, over multiple frames (video footage), this task is highly error-prone. Ref. [55, 56] highlights severe inequity in state-of-the-art, commercial facial analysis systems from companies such as IBM, Microsoft, and Amazon, in a highly controlled setting, for a simple task: gender identification. As late as 2018, the simple task of identifying gender revealed less than 1% error for light-skinned males, compared to 35% error rate for darker-skinned females.

In combination with the documented abuses and unethical applications of facial surveillance systems domestically and internationally, threatening rights of privacy, freedom of expression, freedom of association, and due process [57, 58], these results urge swift action to regulate and possibly prohibit the use of any facial recognition software for automated, online, or lethal decisions.

Table II provides a summary of the key concepts, risks and challenges for LAWS, as described in the the previous sections.

Autonomous weaponry policy

International law: Of all areas of international law, international humanitarian law (IHL) is the most specific and developed, largely out of necessity. Products of the Hague Conference and Geneva Conventions have been able to codify many of the customs that govern conduct during war, as well as laws governing the declaration of war. However, similar to previous times of technological transformation in military technology, international law lags behind the development and use of new classes of weaponry.

In order to better assess the development of international policy on LAWS, past revolutions in military weaponry can provide a framework for analysis. Specifically, emerging nuclear and cyber technologies each created their own *strategic environments*.[1] While a great deal of similarities exist between LAWS and these past technologies in warfare, the incorporation of autonomy into weapon system targeting and engagement can be applied to all existing kinetic systems, rather than creating their own class of weaponry or domain. LAWS are intersectional with these strategic environments rather than a distinct environment of its own. The incorporation of autonomy into existing weapons systems influences how states operate in strategic environments, e.g., conventional, nuclear, and cyber, but it does not create a novel set of conditions effecting security behaviors [59]. As such, as international law continues to adapt to LAWS, policy makers will have to consider how autonomy impacts each strategic environment individually, as well as the impacts of autonomous weapons systems on broader trends in warfare. While the current state of IHL imposes varying structure on existing strategic environments, its applicability to autonomous systems is ambiguous.

United States policy: U.S. Department of Defense Directive 3000.09, *Autonomy in Autonomous Weapons*, defines two echelons of weapons with integrated autonomous systems: autonomous weapons and semi-autonomous weapons. The key differentiating factor between the two in a weapon system's kill chain is in human influence, with autonomous weapons not requiring any human input from target identification to engagement (sometimes called "human out of the loop" systems). The class of semi-autonomous weapons is further broken down into "human *on* the loop systems," where a human can intervene between target identification and engagement, and "human *in* the loop systems," where a human is tasked with selecting or confirming a specific target [1, 60].

This framework provides greater clarity regarding the current state of U.S. policy. According to a report from the Congressional Research Service, the United States does not currently have any autonomous lethal weapons in its inventory. However the same report identified that there are no laws prohibiting the United States from developing or employing LAWS [60]. Further, DOD Directive 3000.09 imposes the ambiguous restriction of "*appropriate* levels of human judgment over the use of force" on weapons to be developed and employed, with no clarification on what is "appropriate". Despite the lack of clarity in official U.S. policy, the U.S. Department of Defense modernization priorities, which include autonomy as well-funded line of effort, shed light on the practical efforts of the U.S. national security apparatus [61, 62].

Policy development challenges: In March of 2019, the United States submitted a report to the UN Convention on Certain Conventional Weapons (CCW) on the application of IHL to the integration of autonomy into weapon systems [63]. The report specifically identified three defined requirements in IHL that are most ambiguous when applied to LAWS [63, 64]: *distinction* between combatants and civilians, *proportionality* of attacks relative to military advantage gained, and *precaution* used in attacks when feasible to reduce risk of civilian casualties.

Since IHL applies to combatants rather than weapon systems, there is a gap in how kill chain decisions are governed for LAWS. In order to provide more insight, the report enumerated three generalized scenarios on autonomy in the employment of a weapon system [63]:

1) An autonomous function of a weapon system could be used to more accurately engage the already-intended target of a commander.
2) An autonomous function could provide information to a commander to inform target selection.
3) An autonomous function of a weapon system could allow for the selection and engagement of targets that were

[1] Strategic environments, as opposed to domains of warfare (e.g., land, air, maritime), are a set of unique "features that condition states' security behaviors" [59], such as the incontestable cost of mutual destruction in nuclear warfare, yielding a strategy of deterrence.

unknown to a commander prior to the function's output.

The CCW report clarifies how the IHL can be applied to LAWS in the above situations. In the first scenario, the target remains consistent, so the use of autonomy is not considered divergent from the use of non-autonomous weaponry. Furthermore, weapons systems with non-autonomous targeting functions in their kill chain are largely not considered within the scope of LAWS. Even so, the justification of the first scenario as within the bounds of IHL assumes autonomous systems are more reliable than humans in the engagement stage of the kill chain. However, autonomous driving has shown this is not always true, with inherent weaknesses and biases in autonomous targeting and tracking [65].

The second scenario describes an autonomous function that is supplemental to the targeting process and can contribute to more informed targeting. In its analysis the report identifies prerequisites for this type of autonomy to be in accordance with IHL: a commander's understanding of the system's accuracy, appropriate synthesizing of other relevant information, and an urgency to make a decision. The first requirement, however, raises questions about practical implementation, as it requires tacticians are sufficiently versed in autonomous systems to understand their accuracy and limitations, and will not fall prey to automation bias in a time when human interpretability of autonomous systems is often sacrificed for greater accuracy [21].

The final scenario posed by the report provides the greatest departure from the current standard practices of warfare. However, in analyzing this scenario, the report draws strong comparison to the use of anti-tank mines, which are used in accordance with IHL without "an express intention at the time of emplacing or activating" [63]. In this comparison, the principles of distinction and proportionality in autonomous weapons are deemed not to diverge significantly enough from conventional equivalents for additional restrictions to be necessary. However, it should be noted the static and predictable nature of anti-tank mines makes for a dubious comparison with loitering munitions which operate over larger geographic and temporal ranges and engage their environment in more complex and potentially unpredictable ways.

As efforts to reconcile existing legal standards to developments in autonomous weaponry continue, implementation remains a central challenge. Any restrictions on LAWS short of a complete ban will need to incorporate validation, enforcement, and accountability processes to assess compliance. These methods raise ambiguity including in how to prevent actors from skirting enforcement and how to integrate accountability methods into the development pipelines of highly complex weapon systems [66].

Future steps: The U.S. government's opposition to increased regulation of autonomous weaponry was laid out in a 2018 white paper to the UN CCW [67]. The argument centered on the potential benefits of autonomous functions, predictable and unforeseeable. While the report specifically cites efforts to stigmatize or ban autonomous weapons systems to be contrary to humanitarian innovative developments, it could be seen to present a false dichotomy.

The report details benefits that might arise from weapons system autonomy, but fails to consider alternatives, short of an outright ban of LAWS, which could reduce collateral damage. Although regulation will likely require a more technically-fluent and fine-tuned approach, the absence of U.S. commitment, or leadership, to create nuanced regulation on the development and employment of LAWS may discourage serious international engagement. A committed effort to regulation would not only avoid the growing risks of LAWS, but it could also encourage more scoped innovation for safer development of lethal autonomous weapons.

Conclusions

The wide and rapid development of lethal autonomous weapons hails a new and perilous era of technological warfare. This work emphasizes the unaccounted risks: the inevitability of development, accessibility, ambiguous attribution or enforcement, and the deployment of unreliable and uninterpretable lethal weapons. The nature and limitations of these weapons systems indicate a high likelihood they will contravene international law, failing to recognize surrendering soldiers or accidentally causing mass civilian casualties. Current policy remains unequipped to handle such risks, and while international pressure grows to limit their usage, commitment to regulation may not acquire sufficient momentum until after more serious catastrophes.

Acknowledgements

The authors thank international experts Daan Kayser, Toby Walsh, and Laura Nolan, as well as our executive editor Kevin McDermott, for their invaluable feedback and guidance.

Citation

Longpre, S., Storm, M., & Shah, R. Lethal autonomous weapons systems & artificial intelligence: trends, challenges, and policies. *MIT Science Policy Review* **3**, 47-56 (2022). https://doi.org/10.38105/spr.360apm5typ.

References

[1] U.S. Department of Defense Directive 3000.09 (2017). Online: https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf,.

[2] Kayser, D. Increasing autonomy in weapons systems: 10 examples that can inform thinking (2021). Online: https://paxforpeace.nl/media/download/Increasing%20autonomy%20in%20weapons%20systems%20-%20FINAL.pdf.

[3] Vick, A. J., Moore, R. M., Pirnie, B. R. & Stillion, J. *Aerospace operations against elusive ground targets* (RAND Corporation, Santa Monica, CA, 2001). https://doi.org/10.7249/MR1398.

[4] Slayton, R. *Arguments that count: physics, computing, and missile defense* (MIT Press, 2013).

[5] Lin-Greenberg, E. Wrestling with killer robots: the benefits and challenges of artificial intelligence for national security. *MIT Schwarzman College of Computing* (2021). https://doi.org/10.21428/2c646de5.1b200d08.

[6] PARS 3 LR anti-tank guided weapon manufacturer datasheet; accessed 2022 (2012). Online: https://www.army-technology.com/projects/lr_trigat/.

[7] Rafael SPIKE anti-tank guided weapon manufacturer datasheet; accessed 2022 (2022). Online: https://www.rafael.co.il/wp-content/uploads/2019/03/Spike-LR2.pdf.

[8] Kane, J. Javelin missile control system patent. *Google Patents* (2000).

[9] Phalanx CWIS U.S. Navy datasheet; accessed 2022 (2022). Online: https://www.navy.mil/Resources/Fact-Files/Display-FactFiles/Article/2167831/mk-15-phalanx-close-in-weapon-system-ciws/.

[10] Vynck, G. D. The U.S. says humans will always be in control of AI weapons. But the age of autonomous war is already here. *The Washington Post* (2021). Online: https://www.washingtonpost.com/technology/2021/07/07/ai-weapons-us-military/.

[11] Gettinger, D. & Michel, A. H. Loitering munitions. *The Center for the Study of the Drone at Bard College* (2017). Online: https://dronecenter.bard.edu/files/2017/02/CSD-Loitering-Munitions.pdf.

[12] United Nations Security Council letter S/2021/229 (8 March 2021). Tech. Rep. Online: https://undocs.org/en/{S}/2021/229.

[13] Sanger, D. E., Schmitt, E., Cooper, H., Barnes, J. E. & Vogel, K. P. U.S. sends thousands of Javelin missiles to Ukraine. *New York Times* (2022). Online: https://www.nytimes.com/2022/03/06/us/politics/us-ukraine-weapons.html.

[14] Pandya, J. The dual-use dilemma of artificial intelligence. *Forbes* (2019). Online: https://www.forbes.com/sites/cognitiveworld/2019/01/07/the-dual-use-dilemma-of-artificial-intelligence/?sh=38c5f58f6cf0.

[15] Lopez, A. M., Comello, J. J. & Cleckner, W. H. Machines, the military, and strategic thought. *Military Review* (2004). Online: http://www.au.af.mil/au/awc/awcgate/milreview/lopez.pdf.

[16] Boulanin, V. & Verbruggen, M. Mapping the development of autonomy in weapon systems. *Stockholm International Peace Research Institute* (2017). Online: https://www.sipri.org/sites/default/files/2017-11/siprireport_mapping_the_development_of_autonomy_in_weapon_systems_1117_1.pdf.

[17] Bieri, M. & Dickow, M. Lethal autonomous weapons systems: future challenges. *CSS Analyses in Security Policy* **164** (2014). Online: https://css.ethz.ch/content/dam/ethz/special-interest/gess/cis/center-for-securities-studies/pdfs/CSSAnalyse164-EN.pdf.

[18] Mouloua, M. & Hancock, P. *Human performance in automated and autonomous systems, current theory and methods* (CRC Press, 2020). https://doi.org/10.7249/MR1398.

[19] Krakovna, V. *et al.* Specification gaming: the flip side of AI ingenuity. *Medium: Deepmind Safety Research* (2020). Online: https://deepmindsafetyresearch.medium.com/specification-gaming-the-flip-side-of-ai-ingenuity-c85bdb0deeb4.

[20] US Department of Defence FY2021 procurement cost estimates (2020). Online: https://www.asafm.army.mil/Portals/72/Documents/BudgetMaterial/2021/Base%20Budget/Procurement/MSLS_FY_2021_PB_Missile_Procurement_Army.pdf.

[21] Lee, J. D. & See, K. A. Trust in automation: designing for appropriate reliance. *Human Factors* **46**, 50–80 (2004). https://doi.org/10.1518/hfes.46.1.50_30392.

[22] Docherty, B. Mind the gap. The lack of accountability for killer robots. *Human Rights Watch* (2015). Online: https://www.hrw.org/report/2015/04/09/mind-gap/lack-accountability-killer-robots#.

[23] Kaag, J. & Kreps, S. *Drone Warfare* (Polity Press, 2014).

[24] Horowitz, M. C. When speed kills: lethal autonomous weapon systems, deterrence and stability. *Journal of Strategic Studies* **42**, 764–788 (2019). https://doi.org/10.1080/01402390.2019.1621174.

[25] Wareham, M. Stopping killer robots: country positions on banning fully autonomous weapons and retaining human control. *Human Rights Watch* (2020). Online: https://www.hrw.org/report/2020/08/10/stopping-killer-robots/country-positions-banning-fully-autonomous-weapons-and.

[26] Autonomous weapons that kill must be banned, insists UN chief (2019). Online: https://news.un.org/en/story/2019/03/1035381.

[27] Brownlee, V. Retaining meaningful human control of weapons systems. *UN Office of Disarmament Affairs* (2018). Online: https://www.un.org/disarmament/update/retaining-meaningful-human-control-of-weapons-systems/.

[28] Autonomous weapons: the ICRC recommends adopting new rules. *International Committee of the Red Cross* (2021). Online: https://www.icrc.org/en/document/autonomous-weapons-icrc-recommends-new-rules.

[29] Sayler, K. International discussions concerning lethal autonomous weapon systems. *Congressional Research Service* (2021). Online: https://sgp.fas.org/crs/weapons/IF11294.pdf.

[30] U.S. Department of Defense. Cluster UAS smart munition for missile deployment (2017). Online: https://www.sbir.gov/sbirsearch/detail/1207935.

[31] Zeitchik, S. The future of warfare could be a lot more grisly than Ukraine. *The Washington Post* (2022). Online: https://www.washingtonpost.com/technology/2022/03/11/autonomous-weapons-geneva-un/.

[32] National Security Commission on Artificial Intelligence. Final report (2020). Online: https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf.

[33] Conger, K. & Metz, C. 'I could solve most of your problems': Eric Schmidt's Pentagon offensive. *New York Times* (2021). Online: https://www.nytimes.com/2020/05/02/technology/eric-schmidt-pentagon-google.html.

[34] Cadena, C. *et al.* Past, present, and future of simultaneous localization and mapping: toward the robust-perception age. *IEEE Transactions on robotics* **32**, 1309–1332 (2016). https://doi.org/10.48550/arXiv.1606.05830.

[35] STM Kargu manufacturer details and operation video; accessed 2022 (2022). Online: https://www.stm.com.tr/en/kargu-autonomous-tactical-multi-rotor-attack-uav".

[36] Luo, W., Yang, B. & Urtasun, R. Fast and furious: real time end-to-end 3D detection, tracking and motion forecasting with a single convolutional net. In *Proceedings of the IEEE conference*

on Computer Vision and Pattern Recognition, 3569–3577 (2018). https://doi.org/10.48550/arXiv.2012.12395.

[37] Casas, S., Sadat, A. & Urtasun, R. MP3: a unified model to map, perceive, predict and plan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14403–14412 (2021). https://doi.org/10.48550/arXiv.2101.06806.

[38] Jung, S., Hwang, S., Shin, H. & Shim, D. H. Perception, guidance, and navigation for indoor autonomous drone racing using deep learning. *IEEE Robotics and Automation Letters* **3**, 2539–2544 (2018). https://doi.org/10.1109/LRA.2018.2808368.

[39] Kawaguchi, K., Kaelbling, L. P. & Bengio, Y. Generalization in deep learning (2017). https://doi.org/10.48550/arXiv.1710.05468.

[40] Alexis, K. Towards a science of resilient robotic autonomy (2020). https://doi.org/10.48550/arXiv.2004.02403.

[41] Scharre, P. *Autonomous weapons and stability*. Ph.D. thesis, King's College London (2020). Online: https://kclpure.kcl.ac.uk/portal/en/theses/autonomous-weapons-and-stability(92cd3d5b-4eb1-4ad5-a9ca-0cce2491e652).html.

[42] Melancon, A.-A. What's wrong with drones? Automatization and target selection. *Small Wars & Insurgencies* **31**, 801–821 (2020). https://doi.org/10.1080/09592318.2020.1743486.

[43] Szegedy, C. *et al.* Intriguing properties of neural networks (2013). https://doi.org/10.48550/arXiv.1312.6199.

[44] Goodfellow, I. J., Shlens, J. & Szegedy, C. Explaining and harnessing adversarial examples (2014). https://doi.org/10.48550/arXiv.1412.6572.

[45] Athalye, A., Carlini, N. & Wagner, D. Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples. In *International conference on machine learning*, 274–283 (PMLR, 2018). https://doi.org/10.48550/arXiv.1802.00420.

[46] Papernot, N. *et al.* The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, 372–387 (IEEE, 2016). https://doi.org/10.1109/EuroSP.2016.36.

[47] Athalye, A., Engstrom, L., Ilyas, A. & Kwok, K. Synthesizing robust adversarial examples. In *International conference on machine learning*, 284–293 (PMLR, 2018). https://doi.org/10.48550/arXiv.1707.07397.

[48] Song, D. *et al.* Physical adversarial examples for object detectors. In *12th USENIX Workshop on Offensive Technologies (WOOT 18)* (2018). https://doi.org/10.48550/arXiv.1807.07769.

[49] Evtimov, I. *et al.* Robust physical-world attacks on machine learning models **2**, 4 (2017). https://doi.org/10.48550/arXiv.1707.08945.

[50] Kurakin, A., Goodfellow, I., Bengio, S. *et al.* Adversarial examples in the physical world (2016). https://doi.org/10.48550/arXiv.1607.02533.

[51] Tu, J. *et al.* Physically realizable adversarial examples for LIDAR object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13716–13725 (2020). https://doi.org/10.48550/arXiv.2004.00543.

[52] Boloor, A., He, X., Gill, C., Vorobeychik, Y. & Zhang, X. Simple physical adversarial examples against end-to-end autonomous driving models. In *2019 IEEE International Conference on Embedded Software and Systems (ICESS)*, 1–7 (IEEE, 2019). https://doi.org/10.48550/arXiv.1903.05157.

[53] Hassaballah, M. & Aly, S. Face recognition: challenges, achievements and future directions. *IET Computer Vision* **9**, 614–626 (2015). https://doi.org/10.1049/iet-cvi.2014.0084.

[54] Phillips, P. J. *et al.* Overview of the face recognition grand challenge. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1, 947–954 (IEEE, 2005). https://doi.org/10.1109/CVPR.2005.268.

[55] Buolamwini, J. & Gebru, T. Gender shades: intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91 (PMLR, 2018). Online: https://proceedings.mlr.press/v81/buolamwini18a.html.

[56] Raji, I. D. & Buolamwini, J. Actionable auditing: investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 429–435 (2019). https://dl.acm.org/doi/10.1145/3306618.3314244.

[57] Najibi, A. Racial discrimination in face recognition technology. *Harvard University* (2020). Online: https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/.

[58] Cyphert, A. B. Reprogramming recidivism: the first step act and algorithmic prediction of risk. *Seton Hall L. Rev.* **51**, 331 (2020). http://dx.doi.org/10.2139/ssrn.3793685.

[59] Fischerkeller, M. Current international law is not an adequate regime for cyberspace. *Lawfare* (2021). Online: https://www.lawfareblog.com/current-international-law-not-adequate-regime-cyberspace.

[60] Sayler, K. Defense primer: U.S. policy on lethal autonomous weapon systems. Tech. Rep. IF11150, Congressional Research Service (2021). Online: https://s3.documentcloud.org/documents/21114192/us-policy-on-lethal-autonomous-weapon-systems-nov-17-2021.pdf.

[61] America's eroding technological advantage: national defense strategy RDT&E priorities in an era of great-power competition with China. Tech. Rep., Govini (2021). Online: https://govini.com/wp-content/uploads/2021/04/Govini_NDS-Priorities-RDTE.pdf.

[62] McBride, C. G-8: CFTs, modernization priorities to see major funding gains in FY-20. *Inside the Army* **30**, 1–7 (2018). https://www.jstor.org/stable/26416583.

[63] United States Government. Implementing international humanitarian law in the use of autonomy in weapon systems. Tech. Rep., United Nations Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects (2019). Online: https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2019)/CCW_GGE.1_2019_WP.5.pdf.

[64] U.S. Department of Defense. Law of war manual (2016). Online https://dod.defense.gov/Portals/1/Documents/pubs/DoD%20Law%20of%20War%20Manual%20-%20June%202015%20Updated%20Dec%202016.pdf?ver=2016-12-13-172036-190.

[65] Koopman, P., Kane, A. & Black, J. Credible autonomy safety argumentation (2018). Online https://users.ece.cmu.edu/~koopman/pubs/Koopman19_SSS_CredibleSafetyArgumentation.pdf.

[66] Mittelsteadt, M. AI verification: mechanisms to ensure AI arms control compliance. Tech. Rep., Center for Security and Emerging Technology (2021). Online: https://cset.georgetown.edu/publication/ai-verification/, https://doi.org/10.51593/20190020.

[67] United States Government. Humanitarian benefits of emerging technologies in the area of lethal autonomous weapon systems. Tech. Rep., United Nations Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, Group of Governmental Experts of the High Contracting Parties (2018). Online: https://ogc.osd.mil/Portals/99/Law%20of%20War/Practice%20Documents/US%20Working%20Paper%20-%20Humanitarian%20benefits%20of%20emerging%20technologies%20in%20the%20area%

20of%20LAWS%20-%20CCW_GGE.1_2018_WP.4_E.pdf?
ver=O0lg6BIxsFt57nrOuz3xHA%3D%3D.