Autonomous Weapon Systems: A Roadmapping Exercise[1]

Ronald Arkin[2], Leslie Kaelbling[3], Stuart Russell[4], Dorsa Sadigh[5], Paul Scharre[6], Bart Selman[7], Toby Walsh[8]

September 9, 2019

Over the past several years, there has been growing awareness and discussion surrounding the possibility of future lethal autonomous weapon systems that could fundamentally alter humanity's relationship with violence in war. Lethal autonomous weapons present a host of legal, ethical, moral, and strategic challenges. At the same time, artificial intelligence (AI) technology could be used in ways that improve compliance with the laws of war and reduce non-combatant harm. Since 2014, states have come together annually at the United Nations to discuss lethal autonomous weapons systems. Additionally, a growing number of individuals and non-governmental organizations have become active in discussions surrounding autonomous weapons, contributing to a rapidly expanding intellectual field working to better understand these issues. While a wide range of regulatory options have been proposed for dealing with the challenge of lethal autonomous weapons, ranging from a preemptive, legally binding international treaty to reinforcing compliance with existing laws of war, there is as yet no international consensus on a way forward.

The lack of an international policy consensus, whether codified in a formal document or otherwise, poses real risks. States could fall victim to a security dilemma in which they deploy untested or unsafe weapons that pose risks to civilians or international stability. Widespread proliferation could enable illicit uses by terrorists, criminals, or rogue states. Alternatively, a lack of guidance on which uses of autonomy are acceptable could stifle valuable research that could reduce the risk of non-combatant harm.

---

[1] This paper grew out of an August 28-29 policy workshop organized by Max Tegmark, Emilia Javorsky and Meia Chita-Tegmark.

[2] School of Interactive Computing, College of Computing, Georgia Institute of Technology
[3] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology
[4] Computer Science Division, University of California, Berkeley
[5] Dept. of Computer Science, Stanford University
[6] Technology & National Security Program, Center for a New American Security
[7] Dept. of Computer Science, Cornell University
[8] School of Computer Science & Engineering, University of New South Wales

International debate thus far has predominantly centered around whether or not states should adopt a preemptive, legally-binding treaty that would ban lethal autonomous weapons before they can be built. Some of the authors of this document have called for such a treaty and would heartily support it, if states were to adopt it. Other authors of this document have argued an overly expansive treaty would foreclose the possibility of using AI to mitigate civilian harm. Options for international action are not binary, however, and there are a range of policy options that states should consider between adopting a comprehensive treaty or doing nothing.

The purpose of this paper is to explore the possibility of a middle road. If a roadmap could garner sufficient stakeholder support to have significant beneficial impact, then what elements could it contain? The exercise whose results are presented below was not to identify recommendations that the authors each prefer individually (the authors hold a broad spectrum of views), but instead to identify those components of a roadmap that the authors are all willing to entertain.[9] We, the authors, invite policymakers to consider these components as they weigh possible actions to address concerns surrounding autonomous weapons.[10]

[9] There is no implication that some authors would not personally support stronger recommendations.

[10] For ease of use, this working paper will frequently shorten "autonomous weapon system" to "autonomous weapon." The terms should be treated as synonymous, with the understanding that "weapon" refers to the entire system: sensor, decision-making element, and munition.

Summary of Issues Surrounding Autonomous Weapons

There are a variety of issues that autonomous weapons raise, which might lend themselves to different approaches. A non-exhaustive list of issues includes:

- The potential for beneficial uses of AI and autonomy that could improve precision and reliability in the use of force and reduce non-combatant harm.
- Uncertainty about the path of future technology and the likelihood of autonomous weapons being used in compliance with the laws of war, or international humanitarian law (IHL), in different settings and on various timelines.
- A desire for some degree of human involvement in the use of force. This has been expressed repeatedly in UN discussions on lethal autonomous weapon systems in different ways.
- Particular risks surrounding lethal autonomous weapons specifically targeting personnel as opposed to vehicles or materiel.
- Risks regarding international stability.
- Risk of proliferation to terrorists, criminals, or rogue states.
- Risk that autonomous systems that have been verified to be acceptable can be made unacceptable through software changes.
- The potential for autonomous weapons to be used as scalable weapons enabling a small number of individuals to inflict very large-scale casualties at low cost, either intentionally or accidentally.

Summary of Components

1. A time-limited moratorium on the development, deployment, transfer, and use of anti-personnel lethal autonomous weapon systems. Such a moratorium could include exceptions for certain classes of weapons.
2. Define guiding principles for human involvement in the use of force.
3. Develop protocols and/or technological means to mitigate the risk of unintentional escalation due to autonomous systems.
4. Develop strategies for preventing proliferation to illicit uses, such as by criminals, terrorists, or rogue states.
5. Conduct research to improve technologies and human-machine systems to reduce non-combatant harm and ensure IHL compliance in the use of future weapons.

Component 1:

States should consider adopting a 5-year, renewable moratorium on the development, deployment, transfer, and use of anti-personnel lethal autonomous weapon systems. Anti-personnel lethal autonomous weapon systems are defined as weapons systems that, once activated, can select and engage dismounted human targets without further intervention by a human operator, possibly excluding systems such as:

- Fixed-point defensive systems with human supervisory control to defend human-occupied bases or installations
- Limited, proportional, automated counter-fire systems that return fire in order to provide immediate, local defense of humans
- Time-limited pursuit deterrent munitions or systems
- Autonomous weapon systems with size above a specified explosive weight limit that select as targets hand-held weapons, such as rifles, machine guns, anti-tank weapons, or man-portable air defense systems, provided there is adequate protection for non-combatants and ensuring IHL compliance[11]

The moratorium would not apply to:

- Anti-vehicle or anti-materiel weapons
- Non-lethal anti-personnel weapons
- Research on ways of improving autonomous weapon technology to reduce non-combatant harm in future anti-personnel lethal autonomous weapon systems
- Weapons that find, track, and engage specific individuals whom a human has decided should be engaged within a limited predetermined period of time and geographic region

[11] The authors are not unanimous about this item because of concerns about ease of repurposing for mass-casualty missions targeting unarmed humans. The purpose of the lower limit on explosive payload weight would be to minimize the risk of such repurposing. There is precedent for using explosive weight limit as a mechanism of delineating between anti-personnel and anti-materiel weapons, such as the 1868 St. Petersburg Declaration Renouncing the Use, in Time of War, of Explosive Projectiles Under 400 Grammes Weight.

Motivation:

This moratorium would pause development and deployment of anti-personnel lethal autonomous weapons systems to allow states to better understand the systemic risks of their use and to perform research that improves their safety, understandability, and effectiveness. Particular objectives could be to:

- ensure that, prior to deployment, anti-personnel lethal autonomous weapons can be used in ways that are equal to or outperform humans in their compliance with IHL (other conditions may also apply prior to deployment being acceptable);
- lay the groundwork for a potentially legally binding diplomatic instrument; and
- decrease the geopolitical pressure on countries to deploy anti-personnel lethal autonomous weapons before they are reliable and well-understood.

Compliance verification:

As part of a moratorium, states could consider various approaches to compliance verification. Potential approaches include:

- Developing an industry cooperation regime analogous to that mandated under the Chemical Weapons Convention, whereby manufacturers must know their customers and report suspicious purchases of significant quantities of items such as fixed-wing drones, quadcopters, and other weaponizable robots.
- Encouraging states to declare inventories of autonomous weapons for the purposes of transparency and confidence-building.
- Facilitating scientific exchanges and military-to-military contacts to increase trust, transparency, and mutual understanding on topics such as compliance verification and safe operation of autonomous systems.
- Designing control systems to require operator identity authentication and unalterable records of operation; enabling post-hoc compliance checks in case of plausible evidence of non-compliant autonomous weapon attacks.
- Relating the quantity of weapons to corresponding capacities for human-in-the-loop operation of those weapons.
- Designing weapons with air-gapped firing authorization circuits that are connected to the remote human operator but not to the on-board automated control system.

- More generally, avoiding weapon designs that enable conversion from compliant to non-compliant categories or missions solely by software updates.
- Designing weapons with formal proofs of relevant properties–e.g., the property that the weapon is unable to initiate an attack without human authorization. Proofs can, in principle, be provided using cryptographic techniques that allow the proofs to be checked by a third party without revealing any details of the underlying software.
- Facilitate access to (non-classified) AI resources (software, data, methods for ensuring safe operation) to all states that remain in compliance and participate in transparency activities.

Component 2: Define and universalize guiding principles for human involvement in the use of force.

- Humans, not machines, are legal and moral agents in military operations.
- It is a human responsibility to ensure that any attack, including one involving autonomous weapons, complies with the laws of war.
- Humans responsible for initiating an attack must have sufficient understanding of the weapons, the targets, the environment and the context for use to determine whether that particular attack is lawful.
- The attack must be bounded in space, time, target class, and means of attack in order for the determination about the lawfulness of that attack to be meaningful.
- Militaries must invest in training, education, doctrine, policies, system design, and human-machine interfaces to ensure that humans remain responsible for attacks.

Component 3: Develop protocols and/or technological means to mitigate the risk of unintentional escalation due to autonomous systems.

Specific potential measures include:

- Developing safe rules for autonomous system behavior when in proximity to adversarial forces to avoid unintentional escalation or signaling. Examples include:
  - No-first-fire policy, so that autonomous weapons do not initiate hostilities without explicit human authorization.
  - A human must always be responsible for providing the mission for an autonomous system.
  - Taking steps to clearly distinguish exercises, patrols, reconnaissance, or other peacetime military operations from attacks in order to limit the possibility of reactions from adversary autonomous systems, such as autonomous air or coastal defenses.
- Developing resilient communications links to ensure recallability of autonomous systems. Additionally, militaries should refrain from jamming others' ability to recall their autonomous systems in order to afford the possibility of human correction in the event of unauthorized behavior.

Component 4: Develop strategies for preventing proliferation to illicit uses, such as by criminals, terrorists, or rogue states:

- Targeted multilateral controls to prevent large-scale sale and transfer of weaponizable robots and related military-specific components for illicit use.
- Employ measures to render weaponizable robots less harmful (e.g., geofencing; hard-wired kill switch; onboard control systems largely implemented in unalterable, non-reprogrammable hardware such as application-specific integrated circuits).

Component 5: Conduct research to improve technologies and human-machine systems to reduce non-combatant harm and ensure IHL-compliance in the use of future weapons, including:

- Strategies to promote human moral engagement in decisions about the use of force;
- Risk assessment for autonomous weapon systems, including the potential for large-scale effects, geopolitical destabilization, accidental escalation, increased instability due to uncertainty about the relative military balance of power, and lowering thresholds to initiating conflict and for violence within conflict;
- Methodologies for ensuring the reliability and security of autonomous weapon systems; and
- New techniques for verification, validation, explainability, characterization of failure conditions, and behavioral specifications.

Definitions used in this document:

Autonomous Weapons System (AWS): A weapon system that, once activated, can select and engage targets without further intervention by a human operator.

Anti-personnel lethal autonomous weapon system: A weapon system that, once activated, can select and engage dismounted human targets with lethal force and without further intervention by a human operator.