Remarks on Lethal Autonomous Weapons Systems to the Tweede Kamer Committee on Foreign Affairs, January 21, 2019 (extended version)

Stuart Russell
University of California, Berkeley

Background

Beginning in 2014, the High Contracting Parties of the Convention on Certain Conventional Weapons (CCW) have held meetings at the United Nations in Geneva to discuss possible limitations on the development and deployment of lethal autonomous weapons systems (AWS). I was invited to address the Informal Meeting of Experts in April, 2015 and the formal Group of Governmental Experts (GGE) in November, 2017.

My remarks will follow the outline from those addresses; they also reflect views expressed in an open letter[1] on July 28, 2015, signed by over 3,700 AI researchers, and in a letter[2] to President Obama written on April 4, 2016, by 41 leading American AI researchers, including almost all of the living presidents of AAAI, the main professional society for artificial intelligence. The UK AI community sent a similar letter to then-Prime Minister David Cameron.

Definitions

The UN defines AWS as having the capacity to "locate, select, and eliminate human targets without human intervention." Some have proposed alternative definitions—for example, the UK Ministry of Defence says that autonomous weapons systems must "understand higher-level intent and direction" and "are not yet in existence and are not likely to be for many years, if at all." Much of the discussion at the UN has been stymied by claims that autonomy is an undefinable and mysterious property that requires learning or even self-awareness. In the view of the AI community, the notion of autonomy is essentially unproblematic in the context of lethal weapons, which is quite distinct from the philosophical context of human autonomy. The autonomy of lethal weapons is no more mysterious than the autonomy of a chess program that decides where to move its pieces and which enemy pieces to eliminate. The key is that the specific targets are not identified and approved, either in advance or at the time of detection, according to human judgment, but are instead selected by the algorithm based on sensory input the algorithm receives after the mission is initiated by a human.

The UN definition also aligns with two key distinctions:

[1] https://futureoflife.org/open-letter-autonomous-weapons/
[2] https://people.eecs.berkeley.edu/~russell/research/LAWS/President-Obama-letter-2016-04-04.pdf

1. Whether the decision to execute a specific lethal attack is made by a human in possession of sufficient information about the individual situation to determine the appropriateness of the attack.
2. Whether the number of attacks that can be carried out can be scaled up independently of the number of humans who are dispatching the weapons.

Feasibility

The feasibility of autonomous weapons is, in my view, not in question, at least for a broad class of missions that might currently be contemplated. All of the component technologies—flight control, swarming, navigation, indoor and outdoor exploration and mapping, obstacle avoidance, detecting and tracking humans, tactical planning, and coordinated attack—have been demonstrated. Building a lethal autonomous weapon, perhaps in the form of a multi-rotor micro-UAV, is easier than building a self-driving car, since the latter is held to a far higher performance standard and must operate without error in a very wide range of complex situations. This is not "science fiction"; autonomous weapons don't have to be humanoid, conscious, and evil; and the capabilities are not "decades away" as claimed by some countries at the CCW.

Legal and humanitarian considerations

UN Special Rapporteur Christof Heyns,[3] Human Rights Watch, the International Committee of the Red Cross,[4] and other experts have expressed concerns about the ability of autonomous weapons to comply with provisions of international humanitarian law (IHL) regarding military necessity, proportionality, and discrimination between combatants and civilians. Discrimination is probably feasible in most situations, even if not perfectly accurate; determining proportionality and necessity is probably not feasible for current AI systems and would have to be established in advance with reasonable certainty by a human operator, for all attacks that the weapons might undertake during a mission. This requirement would therefore limit the scope of missions that could legally be initiated.

Some have raised the issue of accountability. In my view this is not a particular difficulty with autonomous weapons. If the weapon executes an attack that is illegal, then either it could reasonably have been predicted, in which case the operator is directly culpable, or it was unpredictable, in which case the operator is negligent.

Another important component of IHL is the Martens clause, according to which "The human person remains under the protection of the principles of humanity and the dictates of public conscience." In this regard, Germany has stated that it "will not accept that the decision over life and death is taken solely by an autonomous system" while Japan "has

[3] Human Rights Council of the United Nations General Assembly (2013). Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions, Christof Heyns.

[4] International Committee of the Red Cross (2016). Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons.

no plan to develop robots with humans out of the loop, which may be capable of committing murder."[5] BAE Systems, the world's second-largest defense contractor, has asserted that it has no intention of developing autonomous weapons, stating that the removal of the human from the loop is "fundamentally wrong."[6] If the killing of humans by autonomous robots becomes commonplace, it is likely that the "dictates of public conscience" will be very clear in opposing autonomous weapons.

Autonomous weapons as WMDs

Compliance with IHL, even if achievable, is not sufficient to justify proceeding with an arms race in lethal autonomous weapons. Perhaps the most important issue is the effect of AWS on the security of states and their peoples. Here, the message of the AI community, as expressed in the letters mentioned above, is clear: **Because they do not require individual human supervision, autonomous weapons are potentially scalable weapons of mass destruction** (WMDs); essentially unlimited numbers can be launched by a small number of people. This is an inescapable logical consequence of autonomy. As a result, we expect that autonomous weapons will reduce human security at the individual, local, national, and international levels.

I estimate, for example, that roughly one million lethal weapons can be carried in a single container truck or cargo aircraft, perhaps with only 2 or 3 human operators rather than 2 or 3 million. (The Swiss Defense Department constructed an actual prototype based on the recent *Slaughterbots* video, confirming lethality within the specified form factor.[7]) Such weapons would be able to hunt for and eliminate humans in towns and cities, even inside buildings. They would be cheap, effective, unattributable, and easily proliferated once the major powers initiate mass production and the weapons become available on the international arms market. As WMDs, they would have advantages for the victor compared to nuclear weapons or carpet-bombing: they leave property intact and can be applied selectively to eliminate only those who might threaten an occupying force. Finally, whereas the use of nuclear weapons represents a hard threshold that we have not crossed since 1945, there is no such threshold with scalable autonomous weapons. Attacks could escalate smoothly from 100 casualties to 1,000 to 10,000 to 100,000.

Instability and escalation risk

Entrusting a significant portion of a nation's defence capability to autonomous systems is to court instability and risk strategic surprise. For example, the strategic balance between robot-armed countries can change overnight thanks to software updates or cybersecurity penetration, leading to potentially incorrect perceptions of security or strategic superiority. For example, a nation's autonomous weapons might be turned against its own

[5] Statements by the respective ambassadors to the CCW meeting in Geneva, April 2015.

[6] Statement by Sir Roger Carr, BAE chairman, at the World Economic Forum, January 21, 2016; https://www.youtube.com/watch?v=opZR7vLhXVg.

[7] https://youtu.be/9CO6M2HsoIA or autonomousweapons.org.

civilian population. Finally, the possibility of an accidental and rapidly escalating conflict between automated military systems is a serious concern.[8]

A treaty on autonomous weapons

In summary, it seems likely that pursuing an arms race in lethal autonomous weapons would severely reduce international, national, communal, and personal security. The only viable alternative is a treaty that limits the development, deployment, and use of autonomous weapons. Such a treaty would prevent the large-scale manufacturing that would result in wide dissemination of these scalable weapons. Although limiting proliferation of AWS comes with unique challenges, experience with the Chemical Weapons Convention suggests that, with industry cooperation, the residual threat from the diversion of dual-use technology into "home-made" weapons may remain manageable. Moreover, defensive anti-swarm countermeasures could and should remain in place; indeed, their development should be a high priority.

The basic argument that I have presented parallels the argument used by leading biologists to convince President Johnson and then President Nixon to renounce the United States' biological weapons program. This in turn led to the drafting by the United Kingdom of the Biological Weapons Convention and its subsequent adoption.

"Dual-use" technology and civilian AI research

Some parties to the debate have expressed concern that a treaty might impinge on so-called dual-use research that would have civilian benefits. To my knowledge, the AI community does not share these concerns. Biological and chemical weapons are banned, yet biology and chemistry research flourishes. Indeed, a treaty would facilitate the involvement of the AI community on defense-related research, since there would no longer be any danger of the results of AI research being used to create autonomous weapons.

[8] A recent report from the Center for a New American Security, "Autonomous Weapons and Operational Risk," makes many of the same points.