AI and Lethal Autonomous Weapons Systems

Remarks delivered at the meeting of the Group of Governmental Experts on Lethal Autonomous Weapons Systems, Geneva, November 13, 2017

Stuart Russell
Professor of Computer Science, University of California, Berkeley

I will cover a number of topics in a short time: autonomy; the feasibility of autonomous weapons; the possibility of compliance with IHL; the connection between autonomy, scalability, and weapons of mass destruction; and the connection to cyberwarfare.

Much of what I have to say is not just my opinion, but a view shared by a great many in the AI community. This view was expressed in an open letter signed by over 3,000 AI researchers in 2015,[1] and in a letter to the Obama administration written in 2016 by 41 leading American AI researchers, including almost all of the living presidents of AAAI, the main professional society for artificial intelligence.[2,3]

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 *feasibility* of autonomous weapons is also 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. The task of building a lethal autonomous weapon, perhaps in the form of a quadrotor 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.

The issue of compliance with IHL is very important, of course. Discrimination is probably feasible in most situations; 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. Whether the use of lethal autonomous weapons is consistent with the Martens clause is a question that is beyond my expertise, but I note that even the Chairman of BAE Systems, Sir Roger Carr, has stated that he finds such use morally unacceptable.[4]

Compliance with IHL, even if achievable, is not, however, sufficient to justify proceeding with an arms race in lethal autonomous weapons. And let me quote President Obama at this point:

> "I recognize that the potential development of lethal autonomous weapons raises questions that compliance with existing legal norms—if that can be achieved—may not by itself resolve, and that we will need to grapple with more fundamental moral questions about whether and to what extent computer algorithms should be able to take a human life." – Letter dated January 16, 2017.

One must also consider the effect on the security of member states and their peoples. Autonomy necessarily implies *scalability*—the ability of a (small) fixed number of humans to deploy an arbitrarily large number of weapons. I estimate, for example, that 2–3 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. Thus, *autonomous weapons are weapons of mass destruction*—cheap, effective, unattributable, and easily proliferated once the major powers initiate mass production and the weapons become available on the international arms market. I have produced a short film to illustrate this point.[5] With the capability for face recognition, the weapons also provide the possibility of remote, untraceable assassination.

Finally, lethal autonomous weapons allow the extension of cyberwarfare into the physical domain, because software control systems can be infiltrated and modified. For example, a nation's autonomous weapons might be turned against its own civilian population. With no possibility of attribution to an external adversary or individual, one can imagine that the nation's government would be less popular after such an event.

In summary, it seems likely that pursuing an arms race in lethal autonomous weapons would result in a drastic and probably irreversible reduction in international, national, communal, and personal security. The basic argument therefore 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. I think we can all be glad that those steps were taken.

---

[1] https://futureoflife.org/open-letter-autonomous-weapons/
[2] https://people.eecs.berkeley.edu/~russell/research/LAWS/President-Obama-letter-2016-04-04.pdf
[3] https://people.eecs.berkeley.edu/~russell/research/LAWS/President-Obama-reply-2017-01-16.pdf
[4] Statement by Sir Roger Carr, BAE chairman, at the World Economic Forum, January 21, 2016; https://www.youtube.com/watch?v=opZR7vLhXVg.
[5] https://youtu.be/9CO6M2HsoIA or autonomousweapons.org.