James Bret Michael
Associate Editor in Chief

Trustworthiness of Autonomous Machines in Armed Conflict

In September 2017, Russian President Vladimir Putin stated, "Artificial intelligence [AI] is the future, not only for Russia, but for all humankind. It comes with colossal opportunities, but also threats that are difficult to predict. Whoever becomes the leader in this sphere will become the ruler of the world."[1] More recently, a report released by the Harvard Kennedy School's Belfer Center for Science and International Affairs predicted that AI will be just as impactful on national security as other transformative military technologies, such as nuclear, aerospace, cyber, and biotechnology, have been.[2]

One effect of AI on national security has been the development of semiautonomous unarmed and armed defense systems for which a human oversees the operation. An example of such a system is the unmanned surface vessel known as the *Sea Hunter*.[3] The *Sea Hunter*, developed by the Defense Advanced Research Projects Agency, could eventually be armed by the U.S. Navy for conducting antisubmarine, countermine, and other warfare-related operations.

There are even weapons systems envisioned by their stakeholders as one day, in the near future, becoming fully autonomous, such as China's *Marine Lizard* amphibious tank and Russia's armed combat robots.[4,5] Entrepreneurs, such as Elon Musk; diplomats, for example, United Nations Secretary-General António Guterres; and academics, like Prof. Toby Walsh at the University of New South Wales, advocate banning weaponized fully autonomous systems.[6,7] However, given the strategic advantage such systems could give state and nonstate actors to more effectively and efficiently compete in armed conflicts, it will be challenging for some members of the United Nations to agree to an outright ban or otherwise deter the development and use of lethal autonomous systems.

Digital Object Identifier 10.1109/MSEC.2019.2938195
Date of current version: 29 October 2019

Even with an international agreement on banning such weapons systems, it may be difficult to verify that state parties are abiding by the terms of such a treaty. The ability of a system to operate autonomously resides in the system's software, making deception possible: a disguised function in software could be activated to cause the armed system to be switched from non- or semiautonomous to fully autonomous mode, unnoticed by a treaty inspection team. In addition, it is possible to design a plug-and-play fully autonomous system, both the software and hardware, such that the unarmed system could have weaponized modules inserted into it in a just-in-time manner.

My view is that proliferation of weaponized semi- and fully autonomous systems is inevitable. Customary international law and treaties may one day treat such weapons like conventional weapons that have already been banned for use in armed conflict, such as blinding lasers (these are banned under the *Additional Protocol to the 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* "Protocol on Blinding Laser Weapons" [Protocol IV, CCW/CONF. 1/7 (Oct. 12, 1995), hereinafter Protocol IV].[10] Unfortunately, it may turn out that the catalyst for banning weaponized autonomous systems will be the realization of untenable levels of death, injury, and destruction realized from their use in real-world armed conflicts.

Regardless of whether a ban eventually materializes, time is of the essence to determine how to make lethal autonomous systems over their entire lifecycle (i.e., from conception to disposal) as trustworthy as economically and technically feasible. There is even a specific section of the U.S. National AI Research and Development Strategic Plan that specifically calls out the need to address the safety and security of AI-based systems.[8]

Contributors to the content of *IEEE Security & Privacy* already avail themselves of the opportunity to shape policy and law pertaining to systems that incorporate some level of machine intelligence and automatic control. The March–April 2019 issue of *IEEE Security & Privacy* contained several articles regarding the cybersecurity, an important contributor to trustworthiness, of AI-based systems: protecting AI-based systems from cyberattacks, considering the ethics of permitting machines to assume the decision-making tasks of humans, and assessing the behavior of systems that employ complex and opaque AI models.

The idea of autonomous systems is not new. They were used on the battlefield in World War I. It has been a favorite subject of science fiction, such as the 1942 short story "Runaround" in Isaac Asimov's short-story collection *I, Robot*.[9] Asimov explored how robots could interact with humans, and the story includes three laws to be followed by robots.

1. "A robot may not injure a human being or, through inaction, allow a human being to come to harm."
2. "A robot must obey the orders given it by human beings except where such orders would conflict with the First Law."
3. "A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws."

Asimov did not necessarily envision the robots in "Runaround" as being instruments of combat. There are many real-world examples of peaceful applications of robotics and AI (e.g., vehicle automation and search and rescue) in which I would hope that the three laws would be not be violated. Regardless of whether an autonomous system is weaponized, note that systems intended for peaceful uses can be dual use; exploitable security vulnerabilities provide people with malicious intent the opportunity to make the autonomous system do something it was not intended to do, such as violate Asimov's three laws.

I encourage our readers to weigh in on this topic and consider submitting articles to *IEEE Security & Privacy*. Given the potential risks associated with the security vulnerabilities, safety hazards, reliability issues, and so on of autonomous systems and AI-based systems in general, it is important to bring our understanding of cybersecurity and other aspects of trustworthiness, what can be referred to as *technology of the possible*, to bear on influencing *the permissible* (i.e., law) and *the preferable* (i.e., policy or latitude in applying the law).

It has been many years since I was an associate editor in chief of *IEEE Security & Privacy*. I am happy to serve you again in this role. ■

Acknowledgments

The views and conclusions contained herein are those of the author and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotations thereon.

References

1. "'Whoever leads in AI will rule the world': Putin to Russian children on Knowledge Day," *RT News*. Sept. 1, 2017. [Online]. Available: https://www.rt.com/news/401731-ai-rule-world-putin/
2. G. Allen and T. Chan, "Artificial intelligence and national security report," Harvard Kennedy School Belfer Center for Science and International Affairs. July 2017. [Online]. Available: https://www.belfercenter.org/publication/artificial-intelligence-and-national-security
3. "Navy's Sea Hunter drone ship has sailed autonomously to Hawaii and back amid talk of new roles," *The War*

Zone, Feb. 4, 2019. [Online]. Available: https://www.thedrive.com/the-war-zone/26319/usns-sea-hunter-drone-ship-has-sailed-autonomously-to-hawaii-and-back-amid-talk-of-new-roles
4. "Meet the Marine Lizard: Is China's new tank all hype?" *The National Interest*, Apr. 18, 2019. [Online]. Available: https://nationalinterest.org/blog/buzz/meet-marine-lizard-chinas-new-tank-all-hype-53212
5. D. Robitzski, "Russia is planning a 'ground force' of armed military robots," *Futurism*, Mar. 21, 2019. [Online]. Available: https://futurism.com/russia-ground-force-armed-military-robots
6. C. Dreifus, "Toby Walsh, A.I. expert, is racing to stop the killer robots," *NY Times*, July 30, 2019. [Online]. Available: https://www.nytimes.com/2019/07/30/science/autonomous-weapons-artificial-intelligence.html
7. "Autonomous weapons that kill must be banned, insists UN chief," *UN News*, Mar. 25, 2019. [Online]. Available: https://news.un.org/en/story/2019/03/1035381
8. Select Committee on Artificial Intelligence. (2019). The National Artificial Intelligence Research and Development strategic plan: 2019 update. Nat. Sci. Tech. Council. Washington, D.C. [Online]. Available: https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf
9. I. Asimov, "Runaround," in *I, Robot*. New York: Doubleday, 1950, p. 40.
10. Additional Protocol to the 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 (Protocol IV, entitled Protocol on Blinding Laser Weapons), United Nations, Treaty Series, vol. 1380, p. 370; Doc. CCW/CONF.I/16 Part I, Vienna, Oct. 12, 1995. [Online]. Available: https://treaties.un.org/pages/ViewDetails.aspx?src=TREATY&mtdsg_no=XXVI-2-a&chapter=26&lang=en

Digital Object Identifier 10.1109/MSEC.2019.2945375