UNIVERSITY OF CALIFORNIA, BERKELEY

BERKELEY • DAVIS • IRVINE • LOS ANGELES • MERCED • RIVERSIDE • SAN DIEGO • SAN FRANCISCO SANTA BARBARA • SANTA CRUZ

UNIVERSITY OF CALIFORNIA AT BERKELEY

Prof. Stuart Russell
Professor of Computer Science
Michael H. Smith and Lotfi A. Zadeh Chair in Engineering
387 Soda Hall
Berkeley, CA 94720-1776
(510) 642 4964
russell@cs.berkeley.edu

September 28, 2016

To the members of the Defense Innovation Board,

Noting that the upcoming open meeting of the Board intends to discuss "potential application of emerging technologies such as artificial intelligence, autonomy, and man-machine teaming," I would like to offer the following comments. Although I am writing in a personal capacity as an expert in artificial intelligence (AI), the views contained herein are essentially consistent with those expressed in an open letter published on July 28, 2015 and signed by roughly 20,000 scientists and engineers, and with those expressed in a letter written to President Obama on April 4, 2016 and discussed with senior White House staff at a meeting on May 6, 2016. The authors of the letter to President Obama included the majority of senior leaders in the US AI community and 15 members of the National Academies. A similar letter was sent by the UK AI community to Prime Minister Cameron.

My reason for sending these comments is twofold: first, despite the stipulations of DoD Directive 3000.09 requiring "appropriate levels of human judgment over the use of force" and specifically disallowing autonomous selection of human targets even in defensive settings, current and planned DoD research and development and the public comments of some DoD officials suggest that the US is moving towards future deployments of and reliance on lethal autonomous weapons systems (AWS) as a "third offset"; second, various interactions that I and my colleagues around the country have had with DoD officials suggest that there is not a clear understanding at the highest levels of the potential drawbacks of establishing AWS as a primary means of waging war.

Our primary concern is that further movement in this direction is likely to lead to an arms race with negative outcomes for both humanitarian and strategic concerns: in particular, it may lead to a new class of "scalable" weapons of mass destruction – weapons that even small groups could use to attack large populations. Rather than constituting a "third offset" to maintain US military dominance, these developments would instead pose a threat to US and international security.

Legal and humanitarian considerations

UN Special Rapporteur Christof Heyns, Human Rights Watch, the International Committee of the Red Cross, and other experts have expressed concerns about the ability of autonomous weapons to comply with provisions of the laws of armed conflict regarding military necessity, proportionality, and discrimination between combatants and civilians.

Full compliance is probably not feasible at present or in the near future; it requires that machines make subjective and situational judgments that are considerably more difficult than the relatively simple tasks of searching for and engaging potential targets. Even if compliance becomes technically possible, there is of course no guarantee that all parties would use autonomous weapons in legally compliant ways.

Delegating to a machine the decision over the life or death of a human being also raises a fundamental moral question. The Martens Clause of the Geneva Conventions declares that, "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 no plan to develop robots with humans out of the loop, which may be capable of committing murder."[1] 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."[2] At present, the broader public has little awareness of the state of technology and the near-term possibilities, but this will presumably change if the killing of humans by autonomous robots becomes commonplace. At that point, the dictates of public conscience will be very clear but it may be too late to follow them.

Strategic considerations

The component technologies for autonomous weapons, including automated decision making, computer vision, robotics, control systems, and precision manufacturing, have reached the point where fully autonomous weapons are currently feasible for many aerial and naval missions and may soon be feasible for urban warfare. An arms race in autonomous weaponry will lead inevitably to low-cost, mass-produced devices such as flying micro-robots able to hunt for and eliminate humans in towns and cities, even inside buildings. Such devices will form a new, scalable class of weapons of mass destruction with destabilizing properties similar to those of biological weapons. Their scalability is tied intrinsically to their autonomy: once available in large numbers on the arms market, they can be acquired, managed, and launched in the millions with few personnel and almost no infrastructure. Thus, they tip the balance of power away from legitimate states and towards terrorists, criminal organizations, and other non-state actors.

The considerations of the preceding paragraph apply principally to weapons designed for ground warfare and anti-personnel operations, and are less relevant for naval and aerial combat. It is still the case, however, that to entrust a significant portion of our defense capability in any sphere to autonomous systems is to court instability and risk strategic surprise. Autonomous weapons in conflict with other autonomous weapons must adapt their behavior quickly, or else their predictability leads to defeat. This adaptability is

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

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

necessary but makes autonomous weapons intrinsically unpredictable and hence difficult to control. Moreover, 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. Finally, the possibility of an accidental war – a military "flash crash" involving spiraling and unpredictable high-speed interactions among competing algorithms – cannot be discounted.[3] Thus, while there are many ways in which AI and related technologies can contribute to the maintenance of US strategic superiority – e.g., reconnaissance, surveillance, intelligence analysis, tactical and strategic situation assessment, and campaign planning – the development of fully autonomous weapons does not appear to be one of them.

With regard to the obvious question of whether continued adherence to DoD Directive 3000.09 would place the US at a strategic disadvantage: the proper course of action seems to be to design an international treaty that will enforce a ban on lethal 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 these technologies 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-missile systems and anti-robot countermeasures could and should remain in place.

Yours sincerely,

Stuart Russell
Professor of Computer Science, UC Berkeley

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