UNODA Occasional Papers
No. 30, November 2017

Perspectives on Lethal Autonomous Weapon Systems

UNODA
United Nations Office for Disarmament Affairs

United Nations

UNODA

United Nations Office for
Disarmament Affairs

UNODA Occasional Papers

No. 30, November 2017

Perspectives on Lethal Autonomous Weapon Systems

United Nations

The United Nations Office for Disarmament Affairs (UNODA) Occasional Papers are a series of ad hoc publications featuring, in edited form, papers or statements made at meetings, symposiums, seminars, workshops or lectures that deal with topical issues in the field of arms limitation, disarmament and international security. They are intended primarily for those concerned with these matters in Government, civil society and in the academic community.

The views expressed in this publication are those of the authors and do not necessarily reflect those of the United Nations or its Member States.

Symbols of United Nations documents are composed of capital letters combined with figures. These documents are available in the official languages of the United Nations at http://ods.un.org. Specific disarmament-related documents can also be accessed through the disarmament reference collection at https://www.un.org/disarmament/publications/library/.

This publication is available from

www.un.org/disarmament

UNITED NATIONS PUBLICATION
Sales No. E.17.IX.6

ISBN 978-92-1-142324-2
eISBN 978-92-1-362894-2

Printed at the United Nations, New York

A roboticist's perspective on lethal autonomous weapon systems

Ronald C. Arkin
School of Interactive Computing
Georgia Institute of Technology

I. Background on lethal autonomous military robotics

Lethal weapon systems are relatively easy to define. Adding autonomy complicates matters significantly. To a philosopher, autonomy adds moral agency and free will to a robotic system, something that does not yet exist and will not for quite some time, if ever. To a roboticist, however, it simply involves the delegation of decision-making to a machine that has been pre-programmed by a human. This chapter will use the following definition for lethal autonomy:

> The ability to "pull the trigger"—to attack a selected target without human initiation nor confirmation, both in case of target choice or attack command (Foss, 2008).

Note: Portions of this paper have appeared in Arkin, R. C., *Governing Lethal Behavior in Autonomous Systems,* Chapman and Hall Imprint, Taylor and Francis Group, Spring 2009 and are reproduced with permission.

This is restricted only in the same sense as a soldier is restricted: the robot soldier must be given a mission to accomplish and any lethal action must be conducted only in support of that mission. At the highest level, a human is still in the loop, so to speak—commanders must define the mission for the autonomous agent, whether it be a human soldier or a robot. The warfighter, robot or human, must then abide by the rules of engagement and laws of war as prescribed from their training or encoding. Autonomy in this sense is limited when compared to a philosopher's point of view.

Confounding this discussion are those who would delineate levels of autonomy as a basis for discussion. There are many different points of view regarding the terms automation versus autonomy, semi-autonomy, teleautonomy, supervised autonomy, on-the-loop versus in-the-loop, mixed initiative, and on and on. It reached such a level of confusion that a recent defence science board report recommended that none of these terms be used. The specific recommendation was that "the DoD [Department of Defense] should abandon the debate over definitions of levels of autonomy"[1] for a "trade space" approach: a method of analysis of trade-offs over multiple stakeholders and objectives. Here we will not try and map individual systems onto particular levels of autonomy other than to say that all of them involve human involvement to some degree—they are not agents with free will to do whatever they want, and are not systems that are likely to be moral agents anytime soon.

Primary motivators for the use of autonomous, robotic or unmanned systems in the battlefield include the following:

- *Force multiplication.* With robots, fewer soldiers are needed for a given mission and an individual soldier can now do the job that took many before.

[1] Department of Defense, Defense Science Board Task Force Report, "The Role of Autonomy in DOD Systems", July 2012, p. 3.

- *Expanding the battle space.* Robots allow combat to be conducted over larger areas than was previously possible.
- *Extending the warfighter's reach.* Robotics enable an individual soldier to reach deeper into the battle space by, for example, seeing or striking farther.
- *Casualty reduction.* Robots permit removing soldiers from the most dangerous and life-threatening missions.

The initial generation of military robots generally operates under direct human control, such as the "drone" or unmanned aerial vehicles being used by the United States military for air attacks (Singer, 2009; Bergen and Tiedemann, 2009). However, as robotics technology continues to advance, a number of factors are pushing many robotic military systems towards increased autonomy. One factor is that as robotic systems perform a larger and more central role in military operations, there is a need to have them continue to function just as a human soldier would if communication channels are disrupted. In addition, as the complexity and speed of these systems grow, it will be increasingly limiting and problematic for performance levels to have to interject relatively slow human decision-making into the process. As one commentator recently put it, "military systems (including weapons) now on the horizon will be too fast, too small, too numerous, and will create an environment too complex for humans to direct" (Adams, 2002).

Based on these trends, many experts believe that autonomous, and in particular lethal autonomous, robots are an inevitable and imminent development (e.g., Arkin, 2009). Indeed, many military robotic-automation systems already operate at the level where the human is still in charge and responsible for the deployment of lethal force, but not in a directly supervisory manner, as detailed below.[2] Examples

[2] At least 30 nations employ or have in development at least one system of this type, including Australia, Bahrain, Belgium, Canada, Chile, China, Egypt, France, Germany, Greece, India, Israel, Japan, Kuwait, the Netherlands, New Zealand, Norway, Pakistan, Poland, Portugal, Qatar,

generally include close-in weapon systems, anti-submarine weapons, cruise missiles, surface-to-air missiles, fire-and-forget missile systems and anti-personnel and other mines.[3]

These devices are considered to be robotic by most definitions, as they are all capable of sensing their environment and actuating through the application of lethal force.

As early as the end of the First World War, the precursors of autonomous unmanned weapons appeared in a project on unpiloted aircraft conducted by the United States Navy and the Sperry Gyroscope Company (Everett, 2015). Numerous unmanned weaponized robotic systems that employ lethal force and have varying degrees of autonomy are already being developed or are in use.

For a complete listing of weaponized robotic platforms past and present, see Arkin, 2009, chap. 2; Everett, 2015; Roff, 2017; and Human Rights Watch, 2012. A recent United States report stated, "New and powerful robotics systems will be used to perform complex actions, make autonomous systems, deliver lethal force, provide ISR [intelligence, surveillance and reconnaissance] coverage, and speed response times over wider areas of the globe."[4]

II. Ethical autonomy

The development of autonomous, lethal robotics raises questions regarding if and how these systems can adhere to the existing laws of war as well as or better than soldiers. This is

---

the Russian Federation, Saudi Arabia, South Africa, South Korea, Spain, Taiwan, the United Arab Emirates, the United Kingdom and the United States (Scharre and Horowitz, 2015, p. 12).

[3] Antipersonnel mines have been banned by the Ottawa Treaty, although China, the Russian Federation, the United States and 34 other nations are not party to that agreement.

[4] United States Joint Force Development, "Joint Operating Environment 2035: The Joint Force in a Contested and Disordered World", 14 July 2016, p. 17.

no simple task. In the fog of war, it is hard enough for a human to effectively determine whether or not a target is legitimate. Despite the current state of the art, it may be anticipated however that, in the future, autonomous robots may be able to perform better than humans under these conditions for the following reasons:

- The ability to act conservatively; i.e., they do not need to protect themselves in cases of low certainty of target identification. Autonomous, armed robotic vehicles do not need to have self-preservation as a foremost drive, if at all. They can be used in a self-sacrificing manner if needed and without reservation.
- The eventual development and use of a broad range of robotic sensors better equipped for battlefield observations than humans currently possess.
- The absence of emotions, which can cloud human judgment or result in anger and frustration with ongoing battlefield events. In addition, "fear and hysteria are always latent in combat, often real, and they press us toward fearful measures" (Walzer, 1977).
- The avoidance of the human, psychological problem of "scenario fulfillment", a factor believed partly contributing to the downing of an Iranian airliner by the USS *Vincennes* in 1988 (Sagan, 1991). This phenomenon leads to the distortion or neglect of contradictory information in stressful situations, where humans use new incoming information in ways that fit their pre-existing belief patterns, a form of premature cognitive closure. Robots can be developed so that they are not vulnerable to such patterns of behaviour.
- The ability of robots to integrate more information from more sources far faster before responding with lethal force than a human possibly could in real time. These data can arise from multiple remote sensors and intelligence (including human) sources.

- When working in a team of combined human soldiers and autonomous systems as an embedded asset, the potential capability of independently and objectively monitoring ethical behaviour in the battlefield by all parties and reporting infractions that might be observed. This presence alone might possibly lead to a reduction in human ethical infractions.

Considerable research is ongoing in terms of endowing intelligent machines with ethical reasoning or the ability to adhere to moral codes as discussed below (Lin and Bekey, 2014). While "there is every reason to believe that ethically sensitive machines can be created" (Anderson, et al., 2004), there is also widespread acknowledgment regarding the difficulty associated with machine ethics (Moor, 2006; McLaren, 2005 and 2006):

1. Ethical laws, codes, or principles are almost always provided in a highly conceptual, abstract level.
2. Their conditions, premises or clauses are not precise, are subject to interpretation and may have different meanings in different contexts.
3. The actions or conclusions following from the rules are often abstract as well, so, even if the rule is known to apply, the ethically appropriate action may be difficult to execute due to its vagueness.
4. These abstract rules often conflict with each other in specific situations. If more than one rule applies, it is not often clear how to resolve the conflict.

In addition, controversy exists about the correct ethical framework to use in the first place, given the multiplicity of philosophies that exist. In the case of international humanitarian law, the just war theory is agreed upon as the basis for ethical behaviour in the battlefield.

A small sampling of recent and ongoing research on ethical software systems designed to work on autonomous systems is

reviewed below. This is by no means comprehensive but, rather, is intended to provide a snapshot of the current state of the art.

1. Ethical governors

One specific approach has been used in two very different cases for seeking to ensure or guide ethical responses from intelligent robotic systems: the ethical governor. The ethical governor was originally developed as a prototype for use in the application of lethal force in war by an intelligent autonomous robot. It was designed to ensure that these systems comply with international humanitarian law and the rules of engagement—the guidelines for the conduct of warfare. It did so through the application of negative constraints (prohibitions) derived from international humanitarian law and the rules of engagement, ensuring that no laws of war are violated, and the assurance that a positive constraint (obligation) derived from a human commander was present before an attack was permitted. The design and function of this system is well documented elsewhere (Arkin, et al., 2012; Arkin, 2009).

Recently the same underlying approach has been extended to health care—specifically for the management of patient-caregiver relationships in early-stage Parkinson's disease (Shim, et al., 2017). An intervening ethical governor has been designed to help provide a restorative force when this human-human relationship starts to veer beyond acceptable bounds. The intervening ethical governor uses rules derived from occupational therapy manuals, so that a small humanoid robot can intervene when required, as would be the case for a human occupational therapist.

The broad applicability of the ethical governor for enforcing either legal or social norms in a range of applications for autonomous robots should now be apparent. Others such as Welsh (2017) have extended the

concept of the ethical governor using deontic logic, the moral logic of obligations, permissions and prohibitions, to a variety of new domains.

2. Ethical autonomous unmanned undersea vehicles

An example from the United States Naval Postgraduate School involves unmanned undersea vehicles using constraints for "runtime ethics" (Brutzman, et al., 2012 and 2013). Similar to the ethical governor (Arkin, 2009), they use these constraints to monitor the actual execution of the mission for ethical constraint violations before they occur, thus observing the rules of engagement during mission conduct. Their approach entails developing a set of plans using ethical reasoning and then validates them for correctness. Their system is tested in the context of ethical unmanned undersea vehicle search, ensuring that regions that are off-limits to the robot are avoided while still successfully conducting the higher-level mission goals (Davis, et al., 2016).

3. Verifiably ethical autonomous systems

To ensure that ethical behavior is actually obtained, formal verification methods are crucial. Research in the United Kingdom (Dennis, et al., 2013, 2015 and 2016) specifically addresses this area using a Beliefs-Desires-Intentions rational agent architecture with ethical checking to ensure that it selects the most ethical plan available. As in many other pragmatic systems, the ethical principles come from existing rules from society. In this system, the rules are represented in the context of airmanship for unmanned aircraft in civilian aviation, addressing, for example, concerns that arise from low fuel or erratic intruders into common airspace. Their architecture seems readily generalizable to other domains, such as driverless cars and beyond.

4. Case-based ethics for robots

Researchers have investigated using a small humanoid robot to assist in eldercare (Anderson, et al., 2016; Anderson, et al., 2017), using a "case-supported principle-based behaviour paradigm", initially tested only in simulation. The robot identifies the situation it is in, looks at a set of possible actions and then selects the most ethically preferable one (as determined by human ethicists' evaluations a priori). The action predicates are associated with duty satisfaction/violation values, where these duties include rights that serve as guiding principles, such as minimizing harm, respecting autonomy, preventing immobility and the like.

5. Ethical robot architecture

Research in Bristol (Vanderelst and Winfield, 2016) has led to the development of an implemented ethical robot architecture. The system incorporates a discrete ethical layer sitting atop the more traditional robot controller, incorporating a set of ethical rules to determine appropriate courses of action for specific goals. This layer verifies behaviours with respect to ethical performance that are forwarded by the robot controller and can suggest others that are more ethically suitable. Prediction of the consequences of the goals and tasks is then undertaken, followed by evaluation of the predictions, leading to more ethical behaviour than would be achieved otherwise by the robot controller alone. The system was tested on two small humanoid robots to demonstrate an interpretation of Asimov's laws with respect to self-preservation, obedience and human safety. The approach is consequentialist, as it is judged by outcomes rather than inherent duties.

In all these cases, the field of ethical autonomy is still in very early stages of basic research and, although there are hopeful examples that this technology may someday feasibly apply in the battle space, this is likely a decade or two away.

Given the pressing rate of progress in robotics/autonomy as a whole and its rapid penetration in society, it is important that the field move forward post-haste to ensure the safe and ethical deployment of intelligent autonomous robots, especially in the context of armed conflict.

Concurrently, there are major efforts being conducted worldwide aiming to develop policies and standards for the development of these systems. One notable effort is the Institute of Electrical and Electronic Engineers Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems.[5] This strongly interdisciplinary effort and other related ones require worldwide involvement to ensure that the systems we create meet our ethical and societal expectations.

References

Adams, T. (2002). Future Warfare and the Decline of Human Decisionmaking. *Parameters*. U.S. Army War College Quarterly, Winter 2001-02, pp. 57-71.

Anderson, M., S. Anderson and C. Armen (2004). Towards Machine Ethics. *AAAI-04 Workshop on Agent Organizations: Theory and Practice.* San Jose, CA.

Anderson, M., S. Anderson and V. Berenz (2016). Ensuring Ethical Behavior from Autonomous Systems. *Proc. AAAI Workshop: Artificial Intelligence Applied to Assistive Technologies and Smart Environments.* Available from http://www.aaai.org/ocs/index.php/WS/AAAIW16/paper/view/12555.

_____ (2017). A Value Driven Agent: Instantiation of a Case-Supported Principle-Based Behavior Paradigm.

Arkin, R. C. (2009). *Governing Lethal Behavior in Autonomous Robots*. Taylor-Francis.

[5] Available from http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html.

Arkin, R.C., P. Ulam and A. R. Wagner (2012). Moral Decision-making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust and Deception. *Proceedings of the IEEE*, vol. 100, no. 3, pp. 571-589.

Bergen, P. and K. Tiedemann (2009). *Revenge of the Drones: An Analysis of Drone Strikes in Pakistan*. New America Foundation.

Brutzman, D., D. Davis, G. Lucas and R. McGhee (2013). Run-time Ethics Checking for Autonomous Unmanned Vehicles: Developing a Practical Approach. *Proc. 18th International Symposium on Unmanned Untethered Submersible Technology*. Portsmouth, NH.

Brutzman, D., R. McGhee and D. Davis (2012). An implemented universal mission controller with run time ethics checking for autonomous unmanned vehicles—A UUV example. *Autonomous Underwater Vehicles (AUV), 2012 IEEE/OES*. Institute of Electrical and Electronic Engineers.

Davis, D., D. Brutzman, C. Blais and R. McGhee (2016). Ethical mission definition and execution for maritime robotic vehicles: A practical approach. *OCEANS 2016 MTS/IEEE Monterey*, pp. 1-10.

Dennis, Louise, et al. (2013). Ethical choice in unforeseen circumstances. *Conference Towards Autonomous Robotic Systems*. Springer Berlin Heidelberg.

Dennis, Louise A., M. Fisher and A. Winfield (2015). Towards verifiably ethical robot behaviour, arXiv preprint arXiv:1504.03592.

Dennis, Louise, et al. (2016). Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems* 77: 1-14.

Everett, B. (2015). *Unmanned Systems of World Wars I and II.* MIT Press.

Foss, M. (2008). What are Autonomous Weapon Systems and What Ethical Issues do they Raise.

Hawkley, J. (2017). Patriot Wars: Automation and the Patriot Air and Missile Defense System. CNAS Ethical Autonomy Series.

Human Rights Watch (2012). Losing Humanity: The Case Against Killer Robots.

Lin, P., K. Abney and G. Bekey (2014), eds. *Robot Ethics: The Ethical and Social Implications of Robotics.* MIT Press.

Moor, J. (2006). The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems*, July/August, pp. 18-21.

McLaren, B. (2005). Lessons in Machine Ethics from the Perspective of Two Computational Models of Ethical Reasoning. *2005 AAAI Fall Symposium on Machine Ethics*. AAAI Technical Report FS-05-06.

_____ (2006). Computational Models of Ethical Reasoning: Challenges, Initial Steps, and Future Directions. *IEEE Intelligent Systems*, July/August, pp. 29-37.

Roff, H. Dataset: Survey of Autonomous Weapons Systems. Available from https://globalsecurity.asu.edu/robotics-autonomy (accessed on 13 June 2017).

Sagan, S. (1991). Rules of Engagement. *Avoiding War: Problems of Crisis Management*, A. George, ed.. Westview Press.

Scharre, P. and M. Horowitz (2015). An Introduction to Autonomy in Weapon Systems. CNAS Working Paper.

Shim, J., R. C. Arkin and M. Pettinati (2017). An Intervening Ethical Governor for a robot mediator in patient-caregiver relationship: Implementation and Evaluation. *Proc. ICRA 2017*. Singapore.

Singer, P. W. (2009). *Wired for War.* Penguin.

Walzer, M. (1977). *Just and Unjust Wars,* 4th edition. Basic Books.

Welsh, S. (2017). Moral Code: Programming the Ethical Robot, PhD dissertation (draft). University of Canterbury.

Vanderelst, D. and A. Winfield (2016). An Architecture for Ethical Robots, arXiv:1609.02931 (cs.RO).