Dr Ozlem Ulgen – Written evidence (AIC0112)

The ethical implications of developing and using artificial intelligence and robotics in the civilian and military spheres

Summary

Machine-mediated human interaction challenges the philosophical basis of human existence and ethical conduct. Aside from technical challenges of ensuring ethical conduct in artificial intelligence and robotics, there are moral questions about the desirability of replacing human functions and the human mind with such technology. How will artificial intelligence and robotics engage in moral reasoning in order to act ethically? Is there a need for a new set of moral rules? What happens to human interaction when it is mediated by technology? Should such technology be used to end human life? Who bears responsibility for wrongdoing or harmful conduct by artificial intelligence and robotics? This paper seeks to address some ethical issues surrounding the development and use of artificial intelligence and robotics in the civilian and military spheres. It explores the implications of fully autonomous and human-machine rule-generating approaches, the difference between "human will" and "machine will, and between machine logic and human judgment.

About the author

Dr Ozlem Ulgen is a Visiting Fellow at the Lauterpacht Centre for International Law, and Wolfson College, University of Cambridge. She is Senior Lecturer in Law at the School of Law, Birmingham City University, and a barrister (non-practicing) called to the Bar in England and Wales. She specialises in moral and legal philosophy, public international law, and international humanitarian law. Her areas of expertise relate to cosmopolitan ethics in warfare, Kantian ethics and human dignity, and the law and ethics of autonomous weapons. She is currently writing a Routledge-commissioned monograph, *The Law and Ethics of Autonomous Weapons: A Cosmopolitan Perspective*. Forthcoming publications: '*World Community Interest' approach to interim measures on 'robot weapons': revisiting the Nuclear Test Cases* (New Zealand Yearbook of International Law); *Pre-deployment common law duty of care and Article 36 obligations in relation to autonomous weapons: interface between domestic law and international humanitarian law?* (The Military Law and the Law of War Review); *Human dignity in an age of autonomous weapons: are we in danger of losing an 'elementary consideration of humanity'?* (OUP edited collection).

Introduction

1. Artificial intelligence and robotics is pervasive in daily life and set to expand to new levels potentially replacing human decision-making and action. Self-driving cars, home and healthcare robots, and autonomous weapons are some examples. A distinction appears to be emerging between potentially benevolent civilian uses of the technology (e.g. unmanned aerial vehicles delivering medicines), and potentially malevolent military uses (e.g. lethal autonomous weapons killing human combatants). Machine-mediated human interaction challenges the philosophical basis of human existence and ethical conduct. Aside from technical challenges of ensuring ethical conduct in artificial intelligence and robotics, there are moral questions about the desirability of replacing human functions and the human mind with such technology. How will artificial intelligence and robotics engage in moral reasoning in order to act ethically? Is there a need for a new set of moral rules? What happens to human

interaction when it is mediated by technology? Should such technology be used to end human life? Who bears responsibility for wrongdoing or harmful conduct by artificial intelligence and robotics?

2. This paper seeks to address some ethical issues surrounding the development and use of artificial intelligence and robotics in the civilian and military spheres. It explores the implications of fully autonomous and human-machine rule-generating approaches, the difference between "human will" and "machine will, and between machine logic and human judgment.

Fully autonomous and human-machine rule-generating approaches

3. Artificial intelligence and robotics do not possess human rational thinking capacity or a free will to be able to understand what constitutes a rule that is inherently desirable, doable, and valuable for it to be capable of universalisation. But there is human agency in the design, development, testing, and deployment of such technology so that responsibility for implementing moral rules resides with humans. Humans determine which rules are programmed into the technology to ensure ethical use and moral conduct. For these rules to be capable of universalisation they must be "public and shareable".

4. In the civilian sphere, for example, there is much debate about open access and use of artificial intelligence to gather personal data, potentially compromising privacy. In the military sphere, discussions on lethal autonomous weapons under the auspices of the UN Convention on Certain Conventional Weapons represent a process for universalisation of rules which may regulate or ban such weapons. Indeed, there is emerging *opinio juris* among some states for a preventative prohibition rule, and a majority of states recognise that any rules regulating lethal autonomous weapons must take account of ethical, legal, and humanitarian considerations.[1]

5. The potentially broad purposes and uses of artificial intelligence and robotics technology may lead to competing rules emerging which may or may not be capable of universalisation. Some preliminary issues related to the nature and type of rules are considered here.

6. *How* will rules be generated to regulate ethical use and operation of the technology? This depends on whether the technology is intended to completely replace human functions and rational thinking or to complement and supplement such human characteristics. Fully autonomous technology refers to artificial intelligence and robotics replacing human rational thinking capacity and free will so that rules emerge from the technology itself rather than humans. Human-machine integrated technology, on the other hand, refers to technology that supports and assists humans in certain circumstances so that rules are created, influenced, controlled, and tailored by a combination of human and machine interaction and intervention. Both kinds of rule-generating approaches have ethical implications.

a) Fully autonomous rule-generating approach

7. A fully autonomous rule-generating approach would mean the technology produces its own rules and conduct without reference to or intervention from humans. After the initial design

[1] See, O Ulgen, ''World Community Interest' approach to interim measures on 'robot weapons': revisiting the *Nuclear Test Cases*'' (2016) 14 *New Zealand Yearbook of International Law* (forthcoming) Section III.A.

and programming by humans, the technology makes its own decisions. This is "machine learning" or "dynamic learning systems" whereby the machine relies on its own databank and experiences to generate future rules and conduct.[2] Fully autonomous weapons systems, for example, would have independent thinking capacity as regards acquiring, tracking, selecting, and attacking human targets in warfare based on previous experience of military scenarios.[3] Such an approach presents challenges.

8. There is uncertainty and unpredictability in the rules that a fully autonomous weapons system would generate beyond what it has been designed to do, so that it would not comply with international humanitarian law or ethics. In the civilian sphere, fully autonomous technology may generate rules that adversely impact on human self-worth and progress by causing human redundancies, unemployment, and income instability and inequality. Adverse impact on human self-worth and progress, and uncertainty and unpredictability in the rule-generating process are contrary to what is fundamentally beneficial to humankind; such a process cannot produce rules that are inherently desirable, doable, valuable, and capable of universalisation. A perverse "machine subjectivity" or "machine free will" would exist without any constraints.

b) Human-machine rule-generating approach

9. A human-machine rule-generating approach currently exists in both the civilian and military spheres. IBM, for example, prefers the term "augmented intelligence" rather than artificial intelligence because this better reflects their aim to build systems that enhance and scale human expertise and skills rather than replace them.[4] The technology is focused on practical applications that assist people in performing well-defined tasks (e.g. robots that clean houses; robots working with humans in production chains; warehouse robots that take care of the tasks of an entire warehouse; companion robots that entertain, talk, and help elderly people maintain contact with friends, relatives, and doctors). In the military sphere, remotely controlled and semi-autonomous weapons combine human action with weapons technology. Human intervention is necessary to determine when it is appropriate to carry out an attack command or to activate an abort mechanism.

10. This kind of rule-generating approach keeps the human at the centre of decision-making. But what happens if there are interface problems between the human and machine (e.g. errors; performance failures; breakdown of communication; loss of communication link; mis-coordination)?[5] This may prove fatal in human-weapon integrated systems reliant on communication and co-ordination, and a back-up system would need to be in place to suspend or abort operations. What happens if the technology is hacked to produce alternative

---

[2] See, P M Asaro, 'Roberto Cordeschi on Cybernetics and Autonomous Weapons: Reflections and Responses' (2015) 3 *Paradigmi. Rivistadi critica filosofina* 83-107, 96-98; M J Embrechts, F Rossi, F-M Schleif, and J A Lee, 'Advances in artificial neural networks, machine learning,and computational intelligence' (2014) 141 *Neurocomputing* 1-2.

[3] See, *Report of the ICRC Expert Meeting, Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects* (9 May 2014) ('2014 ICRC Report'); *Report of the ICRC Expert Meeting, Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons* (15-16 March 2016) ('2016 ICRC Report'); O Ulgen, 'Autonomous UAV and Removal of Human Central Thinking Activities: Implications for Legitimate Targeting, Proportionality, and Unnecessary Suffering' (forthcoming) 1-45.

[4] F Rossi, 'Artificial Intelligence: Potential Benefits and Ethical Considerations', Briefing Paper to the European Union Parliament Policy Department C: Citizens' Rights and Constitutional Affairs European Parliament (October 2016) <http://www.europarl.europa.eu/RegData/etudes/BRIE/2016/571380/IPOL_BRI(2016)571380_EN.pdf> accessed 26 August 2017.

[5] P M Asaro, 'Roberto Cordeschi on Cybernetics and Autonomous Weapons: Reflections and Responses' (2015) 3 *Paradigmi. Rivistadi critica filosofina* 90-91.

or random rules that cause malfunction, non-performance, or harmful effects? The same problem applies to fully autonomous technology and seems a good reason for restricting use and performance capability to set tasks, controlled scenarios or environments where any potential harm is containable.

Difference between "human will" and "machine will"

11. Kant defined autonomy of will as "the property the will has of being a law to itself (independently of every property belonging to the object of volition)".[6] This may sound chaotic and advocating freedom for humans to do as they please but it is the starting point to explaining how morals come about and how humans should conduct themselves. The ultimate aim of morality is freedom and, therefore, whether conduct is right or wrong is dependent on the extent to which it achieves freedom. If doing something enhances our freedom and can also be universalised to enhance the freedom of others, then it becomes a moral action.

12. Kant's autonomy of will is hard to transpose into technology because it is reliant on concepts such as self-worth, dignity, freedom, and interaction. A machine would not have a sense of these concepts or be able to attach value to them. "Human will" develops through character and experience to inform moral conduct. "Machine learning" or "dynamic learning systems" that generate rules and conduct based on a databank of previous experiences may resemble a form of "machine will" that makes ethical choices based on internally learned rules of behaviour.[7] But the human will is much more dynamic, elusive, and able to cope with spontaneity in reaction to novel situations which sit outside rule-based behavioural action and derive from human experience and intuition.

13. Autonomy of will requires inner and outer development of the person to reach a state of moral standing and be able to engage in moral conduct. This is suggestive of an innate sense of right and wrong. The inner aspect requires adoption and adherence to principles that enhance self-worth and dignity in our person without falling to temptation, personal desires, or external coercion. Examples include avoiding immoral conduct, constantly striving to move from a state of nature to an improved rightful or lawful condition.[8] By enhancing our self-worth and dignity these principles enable us to function freely as rational beings with autonomy of will. The outer aspect is controlled by principles that enable interaction with others and are capable of universalisation. For example, we accept and abide by the general principle that human interaction should be conducted without resorting to violence. In adhering to this principle we are not just motivated by self-preservation but also a higher norm of preserving freedom; if we start conducting our affairs through violence our interaction will become unstable, unpredictable, and unable to guarantee personal freedom or that of others. Can machines emulate this sort of autonomy?

14. Artificial intelligence in autonomous weapons may allow machine logic to develop over time to identify correct and incorrect action, showing a limited sense of autonomy. But the machine

---

[6] I Kant, *The Moral Law: Kant's Groundwork of the Metaphysic of Morals* (H.J. Paton tr, Hutchinson & Co 1969) 101 [440].

[7] M O Riedl, 'Computational Narrative Intelligence: A Human-Centered Goal for Artificial Intelligence' (2016) CHI'16 Workshop on Human-Centered Machine Learning, May 8, 2016, San Jose, California, USA; M O Riedl and B Harrison, 'Using Stories to Teach Human Values to Artificial Agents' (2015) Association for the Advancement of Artificial Intelligence.

[8] I Kant, *The Metaphysics of Morals* (Mary Gregor tr and ed, CUP 1996) 173-218.

does not possess a "will" of its own nor does it understand what freedom is and how to go about attaining it by adopting principles that will develop inner and outer autonomy of will. It has no self-determining capacity that can make choices between varying degrees of right and wrong. The human can decide to question or go against the rules but the machine cannot, except in circumstances of malfunction and mis-programming. It has no conception of freedom and how this could be enhanced for itself as well as humans. The machine will not be burdened by moral dilemmas so the deliberative and reflective part of decision-making (vital for understanding consequences of actions and ensuring proportionate responses) is completely absent.

15. There is a limited sense in which artificial intelligence and robotics may mimic the outer aspect of Kant's autonomy of will. Robots may have a common code of interaction to promote cooperation and avoid conflict among themselves. Autonomous weapons operating in swarms may develop principles that govern how they interact and coordinate action to avoid collision and errors. But these are examples of functional, machine-to-machine interaction that do not extend to human interaction, and so do not represent a form of autonomy of will that is capable of universalisation.

Trust and the technology

16. When we talk about trust in the context of using artificial intelligence and robotics what we actually mean is reliability. Trust relates to claims and actions people make and is not an abstract thing.[9] Machines without autonomy of will, in the Kantian sense, and without an ability to make claims cannot be attributed with trust. Algorithms cannot determine whether something is trustworthy or not. So trust is used metaphorically to denote functional reliability; that the machine performs tasks for the set purpose without error or minimal error that is acceptable.

17. But there is also an extension of this notion of trust connected to human agency in the development and uses to which artificial intelligence and robotics are put. Can we trust the humans involved in developing such technologies that they will do so with ethical considerations in mind (i.e. limiting unnecessary suffering and harm to humans, not violating fundamental human rights)? Once the technology is developed, can we trust those who will make use of it to do so for benevolent rather than malevolent purposes? These questions often surface in debates on data protection and the right to privacy in relation to personal data trawling activities of technologies. Again, this goes back to what values will be installed that reflect ethical conduct and allow the technology to distinguish right from wrong.

The difference between machine logic and human judgment

18. When we compare machines to humans there is a clear difference between the logic of a calculating machine and the wisdom of human judgment.[10] Machines perform cost effective and speedy peripheral processing activities based on quantitative analysis, repetitive actions, and sorting data (e.g. mine clearance; and detection of improvised explosive devices). They are good at automatic reasoning and can outperform humans in such activities. But they lack the deliberative and sentient aspects of human reasoning necessary in human scenarios where artificial intelligence may be used. They do not possess complex cognitive ability to

[9] O O'Neill, *Autonomy and Trust in Bioethics* (CUP 2002).
[10] J Weizenbaum, *Computer Power and Human reason: from judgment to calculation* (1976).

appraise a given situation, exercise judgment, and refrain from taking action or limit harm. Unlike humans who can pull back at the last minute or choose a workable alternative, robots have no instinctive or intuitive ability to do the same. For example, during warfare the use of discretion is important to implementing rules on preventing unnecessary suffering, taking precautionary measures, and assessing proportionality. Such discretion is absent in robots.[11]

19. How will artificial intelligence and robotics engage in moral reasoning in order to act ethically? Should the technology possess universal or particular moral reasoning? Ongoing developments in the civilian and military spheres highlight moral dilemmas and the importance of human moral reasoning to mediate between competing societal interests and values. Companion robots may need to be mindful of privacy and security issues (e.g. protection and disclosure of personal data; strangers who may pose a threat to the person's property, physical and mental well-being) related to assisting their human companion and interacting with third parties (e.g. hospitals; banks; public authorities). Companion robots may need to be designed so that they do not have complete control over their human companion's life which undermines human dignity, autonomy, and privacy. Robots in general may need to lack the ability to deceive and manipulate humans so that human rational thinking and free will remain. Then there is the issue of whether fully autonomous weapons should be developed to replace human combatants in the lethal force decision-making process to kill another human being. Is there a universal moral reasoning that the technology could possess to solve such dilemmas? Or would it have to possess a particular moral reasoning, specific to the technology or scenario?

6 September 2017

[11] See, E Lieblich and E Benvenisti, 'The Obligation to Exercise Discretion in Warfare: Why Autonomous Weapons Systems Are Unlawful', in *Autonomous Weapons Systems Law, Ethics, Policy* (N. Bhuta, S. Beck, R. Geiβ, Liu Hin-Yan, C. Kreβ eds., 2016).