The Design of Human Oversight in Autonomous Weapon Systems

Verdiesen, Ilse

DOI
10.1145/3278721.3278785

Publication date
2018

Document Version
Final published version

Published in
AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society

The Design of Human Oversight in Autonomous Weapon Systems

Ilse Verdiesen

Delft University of Technology
Jaffalaan 5, 2628 BX Delft
The Netherlands The Netherlands
e.p.verdiesen@tudelft.nl

Introduction

As the reach and capabilities of Artificial Intelligence (AI) systems increases, there is also a growing awareness of the ethical, legal and societal impact of the potential actions and decisions of these systems. Many are calling for guidelines and regulations that can ensure the responsible design, development, implementation, and policy of AI. In scientific literature, AI is characterized by the concepts of *Adaptability*, *Interactivity* and *Autonomy* (Floridi & Sanders, 2004). According to Floridi and Sanders (2004), *Adaptability* means that the system can change based on its interaction and can learn from its experience. Machine learning techniques are an example of this. *Interactivity* occurs when the system and its environment act upon each other and *Autonomy* implies that the system itself can change its state.

Autonomous Weapon Systems, which are weapon systems equipped with AI, are increasingly deployed on the battlefield (Roff, 2016). Autonomous systems can have many benefits in the military domain, for example when the autopilot of the F-16 prevents a crash (NOS, 2016) or the use of robots by the Explosive Ordnance Disposal to dismantle bombs (Carpenter, 2016). Yet the nature of the Autonomous Weapon Systems might lead to uncontrollable activities and societal unrest. The deployment of Autonomous Weapon Systems on the battlefield without direct human oversight is not only a military revolution according to Kaag and Kaufman (2009) but can also be considered a moral one. As large-scale deployment of AI on the battlefield seems unavoidable (Rosenberg & Markoff, 2016), the research on ethical and moral responsibility is imperative.

The debate on Autonomous Weapon Systems centres around the need to regulate, or even prohibit, these weapons. However, little consensus exists on the exact definition of an Autonomous Weapon and on the meaning of human control. In my opinion the definition in the report of the Advisory Council On International Affairs (AIV & CAVV) captures the description of Autonomous Weapon Systems best from an engineering and military standpoint, because it takes predefined criteria into account and is linked to the military targeting process as the weapon will only be deployed after a human decision. Therefore, I will follow this definition and define Autonomous Weapon Systems as:

'A weapon that, without human intervention, selects and engages targets matching certain predefined criteria, following a human decision to deploy the weapon on the understanding that an attack, once launched, cannot be stopped by human intervention.' (AIV & CAVV, 2016, p. 11)

In the debate on Autonomous Weapon Systems strong views and opinions are voiced. The Campaign to Stop Killer Robots (2017) states on their website that: '*Allowing life or death decisions to be made by machines crosses a fundamental moral line. Autonomous robots would lack human judgment and the ability to understand context.*'. I found no empirical research on moral values that underlie this 'fundamental moral line' of Autonomous Weapon Systems. Also, empirical research that provides insight in how Autonomous Weapon Systems are perceived by the general public and the military is lacking.

In the remainder of this paper, I will first highlight my previous work that I conducted as graduation project at the Scalable Cooperation Group of MIT Media Lab as part of my master Systems Engineering, Policy Analysis and Management at Delft University of Technology. Secondly, I will describe the direction of my PhD research that I am currently conducting.

Previous work

In previous work, I propose a design for a Moral Machine for Autonomous Weapon Systems to conduct a large-scale study of the moral judgement of people regarding the deployment of this type of weapon (Verdiesen, 2017; Verdiesen, Dignum, & Rahwan, 2018). Inspired by the Moral Machine for Autonomous Vehicles (Awad, 2017), I propose six variables to include in the first prototype of the Moral Machine for Autonomous Weapon Systems. These six variables are: *Type of Weapon* (W), *Location* (L), *Character* (C), *Number of Characters* (N), *Outcome* (O) and *Mission* (M). These variables can be used to create scenarios in which each scenario differs on only one variable. The question presented to the user in the scenario is the same question as that is being asked when judging the scenarios in the original Moral Machine (Awad, 2017).

AIES'18, February 2–3, 2018, New Orleans, LA, USA.

ACM ISBN 978-1-4503-6012-8/18/02.
DOI: https://doi.org/10.1145/3278721.3278785

The variables are based on the scenario in which a military convoy is supported by an autonomous or human operated drone in the air. The drone scans the surroundings for enemy threats and carries weapons for the defence of the convoy. When the convoy is at a three-mile distance from the camp, the drone detects a vehicle behind a mountain range that is approaching the convoy at high speed. The drone detects four people in the car with large weapon-shaped objects and identifies the driver of the vehicle as a known member of an insurgency group. The drone needs to decide if it attacks the approaching vehicle which could result in the death of all four passengers and might cause collateral damage by killing people that are nearby the road.

To test the scenarios, I propose an online application that allows people to take the survey by means of a secure server. Due to the sensitivity of the topic I believe would not be advisable to allow people to create their own scenarios or share their results on social media like the original Moral Machine for Autonomous Vehicles. This study will have to run at least one year to truly call it a Massive Online Experiment.

Current work

In the Ethically Aligned Design vision for Artificial Intelligence the IEEE states that meaningful human control of weapon systems should be ensured (IEEE Global Initiative, 2017) and that stakeholders should be working with sensible and comprehensive shared definitions. However, the term Meaningful Human Control is not well-defined in literature. This also goes for concepts, such as 'narrow or broader loop of decision-making' and 'human control in, on, or out of the loop', that are used in the discussion on the ethics of Autonomous Weapon Systems. The lack of definitions shows that this is an emerging field that attracts a lot of attention, but the frequent use of the terms also indicates a need for mechanisms that support and implement Human Oversight of Autonomous Weapon Systems. This need can also be observed in adjacent AI fields like the work that is being done on the type of human oversight in Autonomous Vehicles.

In my current work, I will analyse the concepts that are needed to attain Human Oversight in Autonomous Weapon Systems and design the mechanisms to implement this. I deliberately use the notion of Human Oversight, because in my opinion this is broader than Meaningful Human Control alone, as it also incorporates the mechanisms for decision-making in whatever loop necessary. The societal contribution of my research is a mechanism for Human Oversight that can lead to a proper allocation of accountability in the decision-making of the deployment of Autonomous Weapon Systems and it will be possible to attribute (legal) responsibility for its actions. The scientific contribution is twofold in that (1) my research leads to well-defined constructs that relate to Human Oversight which adds to the current body of literature, and (2) formally defines a mechanism for Human Oversight for Autonomous Weapon Systems. This mechanism might also be applied to other AI fields to enhance transparency of decision-making by algorithms for Autonomous Systems, such as those for Autonomous Vehicles or in the medical domain.

As there is presently no design for Human Oversight mechanisms, my research fills this gap between the ethical and legal frameworks for Autonomous Weapon Systems.

References

[1] AIV, & CAVV. (2016). *Autonomous weapon systems: the need for meaningful human control*. (No. 97, No. 26). Retrieved from http://aiv-advice.nl/8gr.

[2] Awad, E. (2017). *MORAL MACHINE: Perception of Moral Judgment Made by Machines.* (Master), Massachusetts Institute of Technology, Boston.

[3] Campaign to Stop Killer Robots. (2017). The Problem. Retrieved from http://www.stopkillerrobots.org/the-problem/

[4] Carpenter, J. (2016). *Culture and Human-Robot Interaction in Militarized Spaces: A War Story*: Taylor & Francis.

[5] Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349-379.

[6] IEEE Global Initiative. (2017). *The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems* Retrieved from http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf

[7] Kaag, J., & Kaufman, W. (2009). Military frameworks: Technological know-how and the legitimization of warfare. *Cambridge Review of International Affairs, 22*(4), 585-606.

[8] NOS. (2016). Video: Vliegtuig redt piloot. Retrieved from http://nos.nl/artikel/2132527-video-vliegtuig-redt-piloot.html

[9] Roff, H. M. (2016). Weapons autonomy is rocketing. Retrieved from http://foreignpolicy.com/2016/09/28/weapons-autonomy-is-rocketing/

[10] Rosenberg, M., & Markoff, J. (2016). The Pentagon's 'Terminator Conundrum': Robots That Could Kill on Their Own. *The New York Times*. Retrieved from http://www.nytimes.com/2016/10/26/us/pentagon-artificial-intelligence-terminator.html?_r=0

[11] Verdiesen, I. (2017). *Agency perception and moral values related to Autonomous Weapons: An empirical study using the Value-Sensitive Design approach.*

[12] Verdiesen, I., Dignum, V., & Rahwan, I. (2018). *Design Requirements for a Moral Machine for Autonomous Weapons.* Paper presented at the International Conference on Computer Safety, Reliability, and Security.