The design of human oversight in Autonomous Weapons

Ilse Verdiesen
Delft University of Technology
The Netherlands
e.p.verdiesen@student.tudelft.nl

Master thesis (previous work)

Autonomous Weapons are weapon systems equipped with Artificial Intelligence (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 Weapons 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 Weapons might also lead to uncontrollable activities and societal unrest. The deployment of Autonomous Weapons 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.

In the debate on Autonomous Weapons strong views and opinions are voiced. The Campaign to Stop Killer Robots (2017) states for example 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.*'. We found little empirical research that supports these views or that provide insight in how Autonomous Weapons are perceived by the general public and the military. We also found no empirical research on moral values that underlie the 'fundamental moral line' of Autonomous Weapons. Therefore, the knowledge gap is twofold in that insight is lacking on 1) how Autonomous Weapons are perceived by the military and general public and 2) which moral values people consider important when Autonomous Weapons are deployed in the near future.
The first part of the knowledge gap can be filled by studying the perception of Autonomous Weapons using the agency theory described in the fields of Cognitive Psychology, Artificial Intelligence and Moral Philosophy. The second part of the knowledge gap can be filled by studying known value theories (Beauchamp & Walters, 1999; Friedman & Kahn Jr, 2003; Schwartz, 2012) to see which values people deem important in the deployment of Autonomous Weapons.

In this study, I applied the Value-Sensitive Design (VSD) method as research approach. The VSD is a three-partite approach that allows for considering human values throughout the design process of technology. It is an iterative process for the conceptual, empirical and technological investigation of human values implicated by the design (Davis & Nathan, 2015; Friedman & Kahn Jr, 2003).

The scientific relevance of my study is that I contribute to the academic literature by gaining insight in perception of the general public and the military regarding Autonomous Weapons, and by identifying the moral values the general public and the military relate to Autonomous Weapons. Insight in this is currently lacking and no empirical data on the perception and values related to Autonomous Weapons could be found. By using the Value-Sensitive Design as research approach I show that this method is applicable to structure academic research which could be viewed as case-study for the VSD approach. I also extend the research on the ethical decision-making of Autonomous Vehicles by

Bonnefon, Shariff, and Rahwan (2016) to the domain of Autonomous Weapons.

The societal relevance is that understanding the perception of the general public and military personnel working at the Dutch MOD of Autonomous Weapons, and identifying which moral values they relate to Autonomous Weapons can be used to find common grounds and differences in the debate on this technology initiated by Campaign to Stop Killer Robots (2015) and International Committee for Robot Arms Control (ICRAC). Secondly, the results of this study show how the Value-Sensitive Design method can be applied to Autonomous Weapons to identify the values the military and general public relate to the deployment of these type of weapons. Finally, by identifying the values that are important to incorporate in the design of Autonomous Weapons, the study contributes to a responsible design and deployment of Autonomous Weapons in the future.

Phd proposal

In the first draft of their Ethically Aligned Design vision for Artificial Intelligence (AI) the IEEE states that: '*…meaningful human control of weapons systems is beneficial to society,…*' (IEEE Global Initiative, 2017, p. 68) and that stakeholders should be working with sensible and comprehensive shared definitions. However, in the literature review for my master thesis on the Ethics of Autonomous Weapons I have found that the term meaningful human control is not well-defined. 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 Weapons. On one hand, 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 Weapons. 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, for example on the preferences of people for the Guardian Angel mode versus the Autopilot mode[1].

In my PhD, I will like to analyze the concepts that are needed to attain human oversight in Autonomous Weapons 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 that a mechanism for human oversight would lead to a proper allocation of accountability in the decision-making of the deployment of Autonomous Weapons 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) the mechanism for human oversight for Autonomous Weapons 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 could fill this gap between the ethical and legal frameworks for Autonomous Weapons.

[1] Study done at the Scalable Cooperation research group of MIT Media Lab

References

Beauchamp, T. L., & Walters, L. R. (1999). Contemporary Issues in Bioethics: Wadsworth Pub.

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. Science, 352(6293), 1573-1576.

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

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

Davis, J., & Nathan, L. P. (2015). Value Sensitive Design: Applications, Adaptations, and Critiques. Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains, 11-40.

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

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

Friedman, B., & Kahn Jr, P. H. (2003). Human values, ethics, and design. The human-computer interaction handbook, 1177-1201.

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

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

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

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

Schwartz, S. H. (2012). An overview of the Schwartz theory of basic values. Online readings in Psychology and Culture, 2(1), 11.