The Design of Human Oversight in Autonomous Weapon Systems

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

1 Introduction

Autonomous Weapon Systems (AWS) can be defined as weapons systems equipped with Artificial Intelligence (AI). They are an emerging technology and there is still no internationally agreed upon definition. In my opinion, the definition in the report of the Advisory Council on International Affairs (AIV & CAVV) captures the description of Autonomous Weapons 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 Weapons 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].

AWS are increasingly deployed on the battlefield [Roff, 2016]. In the societal debate on Autonomous Weapon Systems, the concept of Meaningful Human Control (MHC) is often mentioned as requirement [Adams, 2001; Roff & Moyes, 2016; Vignard, 2014], but this term is not-well defined in literature and quantifying the level of control needed is hard [ICRC, 2018]. Some scholars are working on defining the concept of MHC in Autonomous (Weapon) Systems [Ekelhof, 2015; Horowitz & Scharre, 2015; Santoni de Sio & Van den Hoven, 2018]. But in my opinion, MHC will not suffice as requirement to minimize unintended consequences of Autonomous Weapon Systems, because the definition of 'control' implies that you have the power to influence or direct the course of events or the ability to manage a machine (Oxford dictionary). The characteristics *autonomy*, *interactivity* and *adaptability* of AI [Floridi & Sanders, 2004] in Autonomous Weapon Systems inherently imply that control in strict sense in the military domain is not possible, because once a weapon is launched you cannot direct the actions of the weapon (examples are an arrow, missile, torpedo etc.). Therefore, I believe that we will have to take a different approach to minimize unintended consequences of Autonomous Weapons Systems.

Several scholars are describing the concept of Human Oversight in Autonomous Weapon Systems and AI in general. Oversight is defined as 'the action of overseeing something' (Oxford dictionary). HRW and IHRC [2012] state that human oversight on robotic weapons is required to guarantee adequate protection of civilians in armed conflicts and they fear that when humans only retain a limited, or no, oversight role, that they could be fading out the decision-making loop. Just recently Taddeo and Floridi [2018] describe that human oversight procedures are necessary to minimize unintended consequences and to compensate unfair impacts of AI. Nevertheless, current human oversight mechanisms are lacking effectiveness [HRW & IHRC, 2012] and might gradually erode to become meaningless or even impossible [Williams, 2015].

2 PhD Project

In my PhD project, I will analyse the concepts that are needed to attain human oversight in Autonomous Weapon Systems and design a technical architecture to implement this.

2.1 Research Problem

In my research I will build on the method of Bonnemains, Saurel, and Tessier [2018]. New in my approach is that I describe the concept of Human Oversight and to identify, represent and verify the criteria needed for Human Oversight on Autonomous Weapon Systems in order to get insight in the theoretical notion of this concept. To design an architecture for Human Oversight logical rules will be formulized to implement these criteria. These rules will be converted in a logic program and translated to a human readable output that will allow implementation of the architecture. The rules will be validated in scenarios to see if these will actually contribute to the concept of Human Oversight.

2.2 Knowledge Gap

The knowledge gap that I address is that the concept of Human Oversight is not well delineated in literature and I found no architectures for implementing Human Oversight. Therefore, the knowledge gap is twofold in that 1) a theoretical view on the concept of Human Oversight for

Autonomous Weapon Systems and 2) an architecture to implement this concept Human Oversight, are lacking.

2.3 Contributions

If my research is successful, the scientific contribution is twofold in that 1) my research contributes to a well-defined construct of Human Oversight that adds to the current body of literature, and 2) the architecture for Human Oversight for Autonomous Weapon Systems 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. The societal contribution of my research is an architecture for Human Oversight that would lead to a proper allocation of accountability in the decision-making of the deployment of an Autonomous Weapon System and it will be possible to attribute (legal) responsibility for the actions taken by the weapon system by identifying the supervisor of these actions. This will decrease the likelihood of unintended consequences.

2.4 Evaluation

In the validation phase of my research I intend to evaluate the criteria of Human Oversight by running 2 or 3 scenarios to validate the architecture. The type of scenarios could entail both traditional physical weapons systems and cyber weapon systems. At this stage, the validation technique and evaluation metrics need to be determined but could consist of either a simulation, a serious game or a Virtual Reality environment.

2.5 Limitations

The main challenge of my research approach lies in formalizing philosophical definitions in natural language and to translate them in generic computer programmable concepts that can be easily understood and that allows for ethical decisions to be explained. The limitation of my work might be that I am conducting my research in the military domain and my findings might not be generalizable or applicable to other domains as I am studying a very specific field. Also, the formalization of the concept of Human Oversight into rules means that I am interpreting this natural language concept and will lose a lot of context that cannot be captured in logical formulization or computer code.

2.6 Directions for the Remaining Work

I just completed my first year of my PhD at the end of January so there is a lot of remaining work in my PhD project left. I am currently working on the conceptual investigation of the definition of the construct of Human Oversight. Next will be identifying the criteria and formulize these in rules. More generally, directions for remaining work are studying the concept of Human Oversight in related fields, for example in Autonomous Vehicles and in the medical domain to see if my findings are generalizable and applicable to other scientific domains.

References

[Adams, 2001] Thomas. K. Adams. Future warfare and the decline of human decisionmaking. *Parameters, 31*(4), 57-71, 2001.

[AIV & CAVV, 2016] AIV, & CAVV. *Autonomous weapon systems: the need for meaningful human control*. (No. 97, No. 26), 2016. Retrieved from http://aiv-advice.nl/8gr Accessed on: 23 June 2019

[Bonnemains *et al.*, 2018] Vincent Bonnemains, Claire Saurel, & Catherine Tessier. Embedded ethics: some technical and ethical challenges. *Ethics and Information Technology, 20*(1), 41-58, 2018.

[Ekelhof, 2015] Merel Ekelhof. Autonome wapens: een verkenning van het concept Meaningful Human Control. *Militaire Spectator, 184,* 2015.

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

[Horowitz & Scharre, 2015] Michael Horowitz & Paul Scharre. *Meaningful human control in weapon systems: a primer*: Center for a New American Security, 2015.

[HRW & ICRC, 2012] HRW & IHRC. *Losing Humanity: The Case against Killer Robots*, 2012. Retrieved from https://www.hrw.org/sites/default/files/reports/arms1112ForUpload_0_0.pdf Accessed on: 23 June 2019

[ICRC, 2018] ICRC. *Ethics and autonomous weapon systems: An ethical basis for human control?*, 2018. Retrieved from Geneva: https://www.icrc.org/en/download/file/69961/icrc_ethics_and_autonomous_weapon_systems_report_3_april_2018.pdf Accessed on: 23 June 2019

[Roff, 2016] Heather M. Roff. Weapons autonomy is rocketing, 2016. Retrieved from http://foreignpolicy.com/2016/09/28/weapons-autonomy-is-rocketing/ Accessed on: 23 June 2019

[Roff & Moyes, 2016] Heather M. Roff & Richard Moyes. *Meaningful human control, artificial intelligence and autonomous weapons.* Paper presented at the Briefing Paper Prepared for the Informal Meeting of Experts on Lethal Autonomous Weapons Systems, UN Convention on Certain Conventional Weapons, 2016.

[Santoni di Sio Santoni & Van den Hoven, 2018] Filippo Santoni di Sio & Jeroen van den Hoven. Meaningful Human Control over Autonomous Systems: A Philosophical Account. *Frontiers in Robotics and AI, 5*, 15, 2018.

[Taddeo & Floridi, 2018] Mariarosaria Taddeo & Luciano Floridi. How AI can be a force for good. *Science, 361*(6404), 751-752, 2018.

[Vignard, 2014] Kerstin Vignard. The weaponization of increasingly autonomous technologies: considering how meaningful human control might move discussion forward. *UNIDIR Resources, 2,* 2014.

[Williams, 2015] John Williams. Democracy and regulating autonomous weapons: biting the bullet while missing the point? *Global Policy, 6*(3), 179-189, 2015.