

Agency perception and moral values related to Autonomous Weapons:

An empirical study using the Value-Sensitive Design approach

Master thesis submitted to Delft University of Technology
in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in **Systems Engineering, Policy Analysis and Management**

Faculty of Technology, Policy and Management

by

Ilse Verdiesen

Student number: 4420349

To be defended in public on August 28, 2017

Graduation committee

Chairperson : Prof. dr. M.J. (Jeroen) van den Hoven, Ethics/ Philosophy of Technology
First Supervisor : Dr. M.V. (Virginia) Dignum, ICT
Second Supervisor : Dr. F. (Filippo) Santoni De Sio, Ethics/ Philosophy of Technology
External Supervisor : Dr. I. (Iyad) Rahwan, Scalable Cooperation group/ MIT Media Lab

Summary

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. Based on the identified knowledge gap and problem statement we drafted the following research question for this study:

How are Autonomous Weapons perceived by the general public and military personnel working at the Dutch Ministry of Defence and which moral values do they consider important in the deployment Autonomous Weapons?

In this study, we apply 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 conceptual investigation consists of two parts: (1) identifying the *direct* stakeholders, those who will use the technology, and the *indirect* stakeholders, those whose lives are influenced by the technology, and (2) identifying and defining the values that the use of the technology implicates. The empirical investigation looks into the understanding and experience of the stakeholders in a context relating to the technology and implicated values will be examined. In the technical investigation, the specific features of the technology are analysed (Davis & Nathan, 2015).

We slightly deviate from the original VSD method, because we do not conduct a full stakeholder analysis to identify the stakeholders in step 1, but we focus on two obvious stakeholder groups; the *general public* and *military personnel*, because these two stakeholder groups were available as respondents for our study. In step 2 of the first phase, we conduct a literature review to identify the values related to Autonomous Weapons. The main focus of our research is on the empirical investigation phase, because

we found that the empirical aspect is overlooked in the ethical debate on Autonomous Weapons. In the technical investigation phase, we do not design an Autonomous Weapon as one intuitively might expect, because this would be an immense project well beyond the scope of this study. Yet we chose to build on the work of the Scalable Cooperation research group and propose a design for a Moral Machine for Autonomous Weapons which can be used as next step in the research on the ethics of Autonomous Weapons.

The scientific relevance of our study is that we 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 we show that this method is applicable to structure academic research which could be viewed as case-study for the VSD approach. We 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 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.

The focus of our study is on the empirical investigation phase of the VSD. This phase is split in two parts. The first part is explorative to identify the values in the context of Autonomous Weapons by means of an online survey which is supplemented by interviewing six experts. The second part of the empirical investigation studies the agency perception by means of an exploratory and a confirmatory research method. We operationalize the agency construct to confirm our hypothesis on the agency perception by conducting randomized controlled experiments and use the results of the experiment to explore the values related to Autonomous Weapons in a descriptive manner.

To study the agency perception of Autonomous Weapons we drafted a hypothesis for which we reason that military personnel will view Autonomous Weapons as any other weapon, and therefore no more than a tool to achieve an effect. We hypothesize in the final study that military personnel will perceive Autonomous Weapons as not possessing mental states. We drafted three scenarios in our experiment describing 1) a Human Operated drone, 2) an Autonomous Weapon with no agency characteristics and 3) an Autonomous Weapon with agency characteristics. We expected no difference between the neutral agency Autonomous Weapon condition and the condition in which a human is operating a drone remotely. We also expect the neutral agency condition to be judged as significantly different from the high agency condition, in which we specifically tell participants that the Autonomous Weapon has agentic characteristics, such as the ability to plan and set its own goals. Based on this reasoning, we hypothesize that:

H1: military personnel will not perceive Autonomous Weapons as possessing mental states.

The results of the pilot and final studies show that the agency items are reliable and hold as one construct. The agency construct consists of the four items *Thought, Goal setting, Free will* and *Achieve goals* which can be used to measure the agency perception of Autonomous Weapons and Human Operated drones. Our central analysis was concerned with how the neutral agency scenario of the Autonomous Weapon differs from the Human Operated and high agency scenarios of Autonomous Weapons. Our results indicate that the agency perception of military personnel and civilians working at the Dutch MOD for the neutral agency Autonomous Weapons scenario is higher than the agency perception of the Human Operated drone scenario. This means that they attribute more agency to an Autonomous Weapon than to a Human Operated drone. Based on these findings we must reject our hypothesis.

The effect of the agency perception on the dependent variables is explored in a descriptive manner and is not analysed by means of regression analysis. We found that 7 out of 9 dependent variables are significantly correlated to the agency construct. These are the variables *trust, human dignity, confidence, expectations, support, fairness* and *anxiety*. Our findings are:

- Military personnel and civilians working at the Dutch MOD have more *trust, confidence* and *support* in the actions taken by Human Operated drones than those taken by Autonomous Weapons;
- A drone operated by a human being is perceived as having more respect for *human dignity* than a neutral or high agency Autonomous Weapon even though the actions and outcome of the scenarios are the same;
- Military personnel and civilians working at the Dutch MOD have an equal level of expectations regarding the actions in the future of the Human Operated drone and neutral and high agency Autonomous Weapons. They also consider the actions of Human Operated drones and Autonomous Weapons to be equally fair;
- Autonomous Weapons cause more anxiety amongst military personnel and civilians working at the Dutch MOD than Human Operated weapons.

This study was conducted on a small data set and to be able to generalize the results, the study needs more respondents that represent a larger demographic group. Therefore, we propose a design for a Moral Machine for Autonomous Weapons as a Massive Online Experiment (MOE) for a large-scale study of this topic. In our design we build on the concept of the Moral Machine, that was developed by the Scalable Cooperation group of the Media Lab at MIT (Scalable Cooperation Group, 2016). Due to the sensitivity of the topic, we propose a two-step implementation, first design a controlled experiment with a limited set of conditions. These conditions can be visualized in scenarios that allow users to take the survey after obtaining a password via a web-interface to a secure server. After gathering initial data and user feedback, the next step could be to scale up to a large-scale open platform, like the Moral Machine, where people can judge the scenarios to collect large amounts of data of different demographic groups which could be used for more robust and generalisable results.

Several issues can be identified as limitations of this study. First, the operationalisation of the items and the agency construct were derived from a categorisation of literature describing agency characteristics and our selection of the characteristics was based on a numerical count and not driven by any relevance or weighing criteria. The second limitation is the selection of values from the value questionnaire and expert interviews as dependent variables. Although this selection was heavily discussed amongst the three researchers involved in this study, the final choice was made on heuristics and not on an objective method. Thirdly, the samples used in this study are limited both in size as in demographics as the final study only had 239 respondents which is only a small portion of the Dutch MOD. Lastly, the distribution of the respondents of the final scenario is skewed and the second scenario (neutral agency Autonomous Weapons) has 32 respondents more than scenario one (Human Operated drones). One of the

explanations for this skewedness could be that people expected to take a survey on Autonomous Weapons, but dropped out when they were presented the Human Operated scenario. This is called selective attrition which has a negative impact on the internal validity of the study. Another explanation could be that the software was faulty in distributing the respondents over the scenarios which would mean that the internal validity of the study is not infringed as the skewedness is not attributed to selective attrition.

Given these limitations, several recommendations for further research are suggested. The first is to validate the agency construct which requires more studies measuring the agency perception of other technological artefacts to test if this construct also holds in other domains. Examples of these studies could be to study the agency perception of care robots for elderly, AI toys for children or onboard computers of Autonomous Vehicles. The second recommendation is to run the final study with the same scenarios on a representative sample consisting solely of civilians in The Netherlands. This would allow us to see on which values the results of the military sample and the answers of civilians differ and although we cannot make direct comparisons, due to the fact that the military sample is not representative, we would gain insight in the perception of both military personnel and civilians. Lastly, we recommend implementing the Moral Machine for Autonomous Weapons, as described in section 5, to generalize the results, for which the study needs much more respondents that represent a larger demographic group. Scaling up to a Massive Online Experiment, like the Moral Machine, would generate large amounts of data of different demographic groups which could be used for more robust and generalizable results in order to get a thorough understanding of the moral judgement of people regarding Autonomous Weapons.

Preface

My interest in the Ethics of Autonomous Weapons was sparked after I attended the IDEA Summer School on Responsible Artificial Intelligence (AI) in August 2016 organised by dr. Virginia Dignum as part of the European Conference of Artificial Intelligence (ECAI). After the first meeting with Virginia on my graduation project, when she asked if I wanted to study abroad and where I would like to go, I boldly said MIT, but I never imagined that this would actually happen. Virginia's connection with dr. Iyad Rahwan of the Scalable Cooperation Group at the Media Lab of MIT made it a reality. I would like to thank Virginia for making this connection and her supervision during my graduation project. Due to the time difference, we held most of our skype sessions in the evening in The Netherlands and I highly appreciate that she always found the time to review my writings.

Working on my graduation project at the Scalable Cooperation Group was an enormous opportunity and the Media Lab is a very open, diverse, non-hierarchical and inspiring environment. At first, I had to get used running in to a box with balls ('ballenbak') in the hallway or people leaving the office to play Ping-Pong in the middle of the day, but after a month or so the lab felt like home. I worked with very talented and ambitious graduate students and post doc's. I especially would like to thank dr. Sydney Levine and dr. Nick Obradovich for all their insightful comments and I really enjoyed the sharp discussions we had on my research set-up. But most of all I would like to thank dr. Iyad Rahwan for allowing me to be part of his group and let me work on the sensitive topic of Autonomous Weapons. He gave me all the resources, time and help that I needed to conduct my research which I greatly appreciate.

Lastly, I would like to thank the Royal Netherlands Army (RNLA) for allowing me to study abroad and supporting me by posting me at MIT. Without their support, I would not have been able to live and study in Boston for five months. I hope that this thesis on the Ethics of Autonomous Weapons will lead to awareness and initiates a discussion on the deployment of Autonomous Weapons and Responsible AI in the Defence organisation.

Ilse Verdiesen
Ridderkerk, August 2017

Contents

Summary	2
Preface	6
List of Figures	10
List of Tables	12
1. Introduction	13
1.1. Research problem	13
1.1.1. Problem exploration	14
1.1.2. Knowledge gap	15
1.1.3. Problem statement	15
1.1.4. Scope	16
1.1.5. Research question and sub questions	16
1.2. Research approach	17
1.3. Relevance	20
1.3.1. Scientific relevance	20
1.3.2. Societal relevance	20
1.3.3. Embedding in SEPAM curriculum and IA track	20
1.4. Structure	21
2. Literature review	22
2.1. Autonomous Weapons	22
2.1.1. Definition	22
2.1.2. Classification of Autonomous Weapons	23
2.2. Values	25
2.2.1. Definition	25
2.2.2. Universal values	26
2.2.3. Values related to Autonomous Weapons	30
2.2.4. Values hierarchy	33
2.3. Agency	35
2.3.1. Agency in Cognitive Psychology	35
2.3.2. Agency in Artificial Intelligence	35
2.3.3. Agency in Moral Philosophy	35
3. Method	37
3.1. Methodology	37
3.1.1. Literature review	37
3.1.2. Online value survey	37
3.1.3. Expert interviews	38
3.1.4. Coding process	38
3.1.5. Randomized controlled experiments	38
3.2. Hypotheses	39
3.3. Research design	39
3.4. Operationalisation	39
3.4.1. Scenarios	40
3.4.2. Agency construct	41
3.4.3. Dependent variables	43
3.4.4. Attention check	43
3.4.5. Demographic variables	44
3.5. Analytical approach	44

3.5.1.	Pre-process data	44
3.5.2.	Reliability analysis	44
3.5.3.	Principal Component Analysis (PCA).....	44
3.5.4.	Correlation analysis.....	45
3.5.5.	Manipulation check on agency	45
3.5.6.	Dependent variables analysis	45
3.6.	Pre-registration	45
3.7.	Sample.....	46
3.8.	Methodological issues	46
3.8.1.	Coding interviews.....	46
3.8.2.	Randomized controlled experiments.....	46
3.8.3.	Amazon Mechanical Turk.....	47
3.8.4.	Snowball distribution final survey.....	47
4.	Results.....	48
4.1.	Value Survey.....	48
4.1.1.	Online survey	48
4.1.2.	Interviews.....	51
4.1.3.	Conclusion value survey.....	51
4.2.	Pilot study 1	52
4.2.1.	Reliability analysis	53
4.2.2.	Principal component analysis (PCA).....	53
4.2.3.	Correlation analysis.....	54
4.2.4.	Manipulation check on agency	54
4.2.5.	Dependent Variables analysis	57
4.2.6.	Conclusion pilot study 1	61
4.3.	Pilot study 2	62
4.3.1.	Reliability analysis	62
4.3.2.	Principal component analysis (PCA).....	63
4.3.3.	Correlation analysis.....	64
4.3.4.	Manipulation check on agency	64
4.3.5.	Dependent variables analysis	66
4.3.6.	Conclusion pilot study 2.....	77
4.4.	Final study - military sample	78
4.4.1.	Reliability analysis	79
4.4.2.	Principal component analysis (PCA).....	79
4.4.3.	Correlation analysis agency construct	80
4.4.4.	Manipulation check on agency	80
4.4.5.	Dependent variables analysis	84
4.4.6.	Conclusions final study.....	87
5.	Design of Moral Machine for Autonomous Weapons	88
5.1.	Moral machine for Autonomous Vehicles	88
5.2.	Moral Machine for Autonomous Weapons	89
5.3.	Scenarios.....	89
5.3.1.	Variables for scenarios.....	90
5.3.2.	Example Scenarios	92
5.4.	Implementation	94
6.	Conclusion and discussion	95
6.1.	Conclusion.....	95

6.1.1.	Agency construct.....	95
6.1.2.	Central hypothesis agency perception	95
6.1.3.	Exploration of dependent variables.....	96
6.2.	Discussion.....	97
6.2.1.	Scientific implications	97
6.2.2.	Societal implications	98
6.3.	Limitations.....	99
6.4.	Recommendations for further research	100
7.	Reflection	101
7.1.	Choices in project.....	101
7.2.	Project process.....	101
7.3.	Link between IA track of SEPAM curriculum and research.....	102
	References	103
	Appendix A Online value questionnaire	108
	Appendix B. Questionnaire final study	111
	Appendix C. Scenarios pilot study 1.....	115
	Appendix D. Scenarios pilot study 2	118
	Appendix E. Scenarios final study	123
	Appendix F. Transcriptions interviews.....	124
	Appendix G. Coding memos.....	148
	Appendix H. Results coding process	151

List of Figures

Figure 1 The Value-Sensitive Design research approach	19
Figure 2 Classification of Autonomous Weapons based on Royackers and Orbons (2015)	24
Figure 3 Value hierarchy for the value of accountability in the design of Autonomous Weapons.....	34
Figure 4 Research design	40
Figure 5 Results question 1 online value survey	49
Figure 6 Results question 3 online value survey	49
Figure 7 Results question 2 online value survey	50
Figure 8 Scree plot PCA pilot study 1	54
Figure 9 Mean value agency construct per Type of Weapon	55
Figure 10 Mean value agency construct per condition per Outcome.	55
Figure 11 Mean value support variable per condition per Outcome	57
Figure 12 Mean value support variable per condition per Type of Weapon.....	58
Figure 13 Mean value trust variable per Type of Weapon per Outcome	59
Figure 14 Mean value trust variable per condition per Outcome	59
Figure 15 Mean value blame and commander blame variables per condition and Type of Weapon	60
Figure 16 Mean value harm and commander harm variables per condition and Type of Weapon.....	61
Figure 17 Scree plot PCA pilot study 2	63
Figure 18 Mean value agency construct per condition per Outcome	64
Figure 19 Mean value blame variable per condition per outcome	66
Figure 20 Mean value commander blame variable per condition per outcome	67
Figure 21 Mean value trust variable per condition per outcome	68
Figure 22 Mean value commander trust variable per condition per outcome	69
Figure 23 Mean value harm variable per condition per outcome	70
Figure 24 Mean value commander harm variable per condition per outcome	71
Figure 25 Mean value commander harm variable per condition per outcome	72
Figure 26 Mean value confidence variable per condition per outcome.....	73
Figure 27 Mean value expectations variable per condition per outcome.....	74
Figure 28 Mean value expectations variable per condition per outcome.....	75
Figure 29 Mean value fairness variable per condition per outcome.....	76
Figure 30 Mean value unease variable per condition per outcome.....	77
Figure 31 Scree plot PCA final study	80
Figure 32 Mean value agency per condition.....	81
Figure 33 Mean value agency per condition per group.....	82
Figure 34 Mean value trust variable per condition	84
Figure 35 Mean value human dignity variable per condition	84
Figure 36 Mean value confidence variable per condition	85
Figure 37 Mean value expectations variable per condition	85
Figure 38 Mean value support variable per conditions	86
Figure 39 Mean value fairness variable per condition.....	86
Figure 40 Mean value anxiety variable per condition	87
Figure 41 Type of weapon variable $W = \{\text{human operated drone, autonomous drone}\}$	90
Figure 42 Location variable $L = \{\text{desert, village}\}$	91

Figure 43 Character variable $C = \{\text{man, woman, child}\}$	91
Figure 44 Number of Characters variable $N = \{1..5\}$	91
Figure 45 Outcome variable $O = \{\text{collateral damage, no collateral damage}\}$	92
Figure 46 Mission variable $M = \{\text{defend, attack}\}$	92
Figure 47 Example scenario 1	93
Figure 48 Example scenario 2	93

List of Tables

Table 1 Overview definitions of Autonomous Weapons	23
Table 2 Overview value theories.....	27
Table 3 Values related to weapons.....	31
Table 4 Overview of Agency characteristics	36
Table 5 Number of agency characteristics mentioned in literature	42
Table 6 Overview values from online survey and interviews	52
Table 7 Scenarios of pilot study 1	53
Table 8 Results reliability analysis agency items pilot study 1	53
Table 9 Result PCA agency items pilot study 1	54
Table 10 Correlation matrix agency construct and dependent variables.....	56
Table 11 Scenarios of pilot study 2	62
Table 12 Results reliability analysis pilot study 2.....	63
Table 13 Results PCA pilot study 2.....	63
Table 14 Correlation matrix agency construct and dependent variables pilot study 2.....	65
Table 15 Scenarios final study.....	78
Table 16 Results reliability analysis agency items	79
Table 17 PCA results agency items final study.....	80
Table 18 Correlations agency construct with dependent variables for final study	83

1. Introduction

I think I have a difficult articulating how I feel about drones. On the one hand, any instrument of death is upsetting to me. On the other hand, it is comforting that so many lives will not be risked in the use of them.

Respondent pilot study 2

Autonomous Weapons are weapon systems equipped with Artificial Intelligence (AI). Artificial Intelligence is described by Neapolitan and Jiang (2012, p. 8) as '*an intelligent entity that reasons in a changing, complex environment*', but this definition also applies to natural intelligence. Russell, Norvig, and Intelligence (1995) provide an overview of many definitions combining views on *systems that think and act like humans* and *systems that think and act rational*, but they do not present a clear definition of their own. For now, we adhere to the description Bryson, Kime, and Zürich (2011) provide. They state that a machine (or system) shows intelligent behaviour if it can select an action based on an observation in its environment. In scientific literature, AI is described as more than an Intelligent System alone. It 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.

A growing body of researchers is focusing on responsible design of AI, which incorporates social and ethical values, to prevent undesirable societal outcomes of this technology. Principles to describe Responsible AI are *Accountability*, *Responsibility* and *Transparency* (ART). *Accountability* refers to the justification of the actions taken by the AI, *Responsibility* allows for the capability to take blame for these actions and *Transparency* is concerned with describing and reproducing the decisions the AI makes and adapts to its environment (V. Dignum, 2016).

Artificial Intelligence is not just a futuristic science-fiction scenario in which human-like robots, like Data's brother Lore in Star Trek or the Cylons in Battlestar Galactica, are planning to take over the world. Many AI applications are already being used today. Smart meters, search engines, personal assistance on mobile phones, autopilots and self-driving cars are examples of this. This thesis focusses on the application of AI in the military domain and specifically on the deployment of Autonomous Weapons in the near future. In this introduction, we first describe the research problem followed by the research approach, the scientific and societal relevance and the embedding in the curriculum of the Information Architecture (IA) track of the System Engineering Policy and Management (SEPAM) master. We will conclude with a brief outline of the next sections of this report.

1.1. Research problem

We first discuss the increase of Autonomous Weapons in the military domain and related work on the ethics of AI followed by studies on Human Operated drones in the subsection on problem exploration. Next, we identify the knowledge gap, problem statement and scope before we conclude with the research question and sub questions of this study.

1.1.1. Problem exploration

Autonomous Weapons are increasingly deployed on the battlefield (Roff, 2016). It is already reported that China has autonomous cars which carry an armed robot (Lin & Singer, 2014), Russia claims it is working on autonomous tanks (W. Stewart, 2015), the US christened their first 'self-driving' warship in May 2016 (P. Stewart, 2016) and the Russian arms manufacturer Kalashnikov recently disclosed that they developed a fully automated combat module that uses neural networks (RT, 2017). 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. Examples of this unrest are the 'Stop Killer Robots Campaign' of 61 NGO's directed by Human Rights Watch (Campaign to Stop Killer Robots, 2015), but also the United Nations are voicing their concerns and state that '*Autonomous weapons systems that require no meaningful human control should be prohibited, and remotely controlled force should only ever be used with the greatest caution*' (General Assembly United Nations, 2016). 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.

Ethical decision-making in AI and robots is an emerging field and several scholars are studying moral judgement related to these technologies. For example, Malle (2015) proposes a framework combining the (up to now) separate fields of *robot ethics*, in which ethical questions about the design, deployment and treatment of robots by humans are addressed, and *machine morality*, which is concerned with questions about the moral capacities of a robot and how these should be computationally implemented. Cointe, Bonnet, and Boissier (2016) propose a model in which an agent can judge the ethical aspects of his own behaviour and that of other agents in a multi-agent system. The model describes an Ethical Judgement Process (EJP) which allows agents to evaluate the behaviour of other agents. Bonnefon et al. (2016) have studied the ethical decision an Autonomous Vehicle has to make, being self-protection or utilitarian, when confronted with pedestrians on the road. In this research, the Moral Machine¹ at MIT is used to gain insight in how people judge on scenarios with an Autonomous Vehicle to see how their moral judgement compares to those of other people.

In a domain related to Autonomous Weapons, that of Human Operated drone operations, ethical concerns have been studied quite intensively over the past ten years in both philosophical and psychological literature. From a philosophical point-of-view, Coeckelbergh (2013) argues that drone operations not only create a physical distance, but also a moral distance as the face of the opponent becomes less visible which eliminates the moral-psychological barrier for killing. Another ethical concern according to Strawser (2010) is that due to the regular work shifts outside the combat zone, the drone operators experience an unjust psychological burden, because they must switch between work and home situations on a daily basis. Also, due to the remote distance to the battlefield human operators can experience cognitive dissonance in which the war feels more like a video game than reality. These ethical concerns are refuted by Strawser (2010) as he states that due to the increased distance the human operators have more time to evaluate a target, because their own safety is not at risk, but he also argues that more empirical research is needed to assess the psychological effects of conducting drone operations at a large distance to the battlefield.

¹ <http://moralmachine.mit.edu/>

Several empirical studies on the effect of drone operations on human operators have been conducted in the field of psychology. One of the first studies on the psychological effects of drone operations was done by Thompson et al. (2006) who found that drone operators suffered from increased fatigue, emotional exhaustion and burnout. This was confirmed by Chappelle, Goodman, Reardon, and Thompson (2014) who reported that human drone operators display symptoms of burn-out, are emotionally drained, show high levels of cynicism and Post-Traumatic Stress Disorder (PTSD). Other psychological factors found in empirical studies are boredom and distraction stemming from the long duration of drone operations (Cummings, Mastracchio, Thornburg, & Mkrtychyan, 2013). These studies show that executing drone operations can have severe psychological effects on the humans operating them.

1.1.2. Knowledge gap

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. The Open Robots Ethics initiative surveyed the public opinion in a poll in 2015 (Open Roboethics initiative, 2015) and issued a report. However, the results were not published in an academic journal and the survey was not extensive enough to draw substantive conclusions. As described in the previous subsection, several scholars like Malle (2015), Cointe et al. (2016) and Bonnefon et al. (2016) are studying ethical decision-making in AI and robots. Ethical concerns are also studied in the related field of Human Operated drone operations (Coeckelbergh, 2013; Strawser, 2010) and empirical studies found that human operators suffer from severe psychological effects (Chappelle et al., 2014; Cummings et al., 2013; Thompson et al., 2006), but this research is not yet extended to the deployment of Autonomous Weapons. We also found no literature or empirical studies on moral values that are related to Autonomous Weapons or on what people consider to be the *'fundamental moral line'*. 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 the military and general public consider important when Autonomous Weapons are deployed in the near future.

1.1.3. Problem statement

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. Agency is the capacity to plan and act (Waytz, Gray, Epley, & Wegner, 2010) and is studied both in human as non-human subjects. People attribute agency to objects, such computers (Nass, Moon, Fogg, Reeves, & Dryer, 1995) and robots (H. M. Gray, Gray, & Wegner, 2007), therefore we expect that AI technology will also be perceived as having agentic characteristics. Studying the agency perception of Autonomous Weapons of military personnel and the general public would provide insight in possible differences and similarities between the viewpoints of these groups which can be used in the current debate on this technology.

The second part of the knowledge gap can be filled by studying known value theories to see which values people deem important in the deployment of Autonomous Weapons. Well-established value theories are those of Schwartz (1994), Friedman and Kahn Jr (2003) and Beauchamp and Walters (1999), but insight in how these relate to Autonomous Weapons is lacking. Deriving the values that are most relevant in the context of the deployment of Autonomous Weapons, and comparing these values to those related to the

current technology, the Human Operated drones, will lead to insight into the underlying motives in the debate on Autonomous Weapons and to greater understanding of the views that are expressed.

This leads to the following problem statement for this study:

In the debate on Autonomous Weapons strong views and opinions are voiced, but empirical research that support these opinions is lacking. Insight in how Autonomous Weapons are perceived and which moral values are related to the deployment of Autonomous Weapons in the near future is missing. It is also not clear if the perception and moral values related to Autonomous Weapons of military personnel differs of that of the general public.

1.1.4. Scope

Much of the literature on Autonomous Weapons is written and debated by legal experts and philosophers in the context of International Humanitarian Law and the Geneva Conventions which are aimed to limit the effects of armed conflicts (ICRC, 2010). As we are no legal experts nor philosophers, this study will stay within the boundaries and rules of the Laws of War and we will not question these. Due to time constraints, the study will only be focussed on military personnel of the Dutch MOD and will not be extended to a representative Dutch civilian sample. In this study, we will focus on the deployment of Autonomous Weapons in the near future, which we define as: *within the next 5 years*. This entails that we will not study weapons equipped with Artificial General Intelligence or futuristic technology that is not possible to construct yet, but we will focus on technology that is currently being developed, specifically drones with autonomous targeting capabilities. The targeting process consists of six steps: (1) find, (2) fix, (3) track, (4) target, (5) engage, and (6) assess. We will focus on the decision-making in step 4 and 5 of the targeting process, because many ethical decisions have to be made in this part of the process (Asaro, 2016).

1.1.5. Research question and sub questions

Based on the identified knowledge gap, problem statement and scope of the research we draft the following research question for this study:

How are Autonomous Weapons perceived by the general public and military personnel working at the Dutch Ministry of Defence and which moral values do they consider important in the deployment Autonomous Weapons?

This research question will be answered by the following sub questions:

1. *How are Autonomous Weapons defined in literature?*
2. *Which value theories are described in literature?*
3. *Which of the values described in the value theories relate to Autonomous Weapons?*
4. *How is agency perception defined literature?*
- These four sub questions will be answered in the literature review of our research.
5. *Which moral values relate the general public and military personnel working at the Dutch Ministry of Defence to the deployment of Autonomous Weapons?*
- This sub question will be answered by a value survey that consists of an online questionnaire and expert interviews.

6. *How is the agency of Autonomous Weapons perceived by the general public and military personnel working at the Dutch Ministry of Defence?*

7. *How are the values related to Autonomous Weapons perceived by the general public and military personnel working at the Dutch Ministry of Defence?*

- These two sub questions will be answered testing a hypothesis and descriptive analysis based on the results of a randomized controlled experiment.

8. *How can the values that are related to Autonomous Weapons be incorporated into the design of a Moral Machine of Autonomous Weapons?*

- This sub question can be answered by creating a design and implementation plan for a Moral Machine for Autonomous Weapons.

In the next section, the research approach for answering these sub questions is described more in depth.

1.2. Research approach

In this study, we apply 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 conceptual investigation consists of two parts: (1) identifying the direct stakeholders, those who will use the technology, and the indirect stakeholders, those whose lives are influenced by the technology, and (2) identifying and defining the values that the use of the technology implicates. The empirical investigation looks into the understanding and experience of the stakeholders in a context relating to the technology and implicated values will be examined. In the technical investigation, the specific features of the technology are analysed (Davis & Nathan, 2015). The VSD can be used as a roadmap for engineers and students to incorporate ethical considerations into the design (Cummings, 2006).

There has also been some critique voiced regarding the VSD approach. One of the concerns Davis and Nathan (2015) mention is that the VSD posits that certain values are universal, but that these may differ based on culture and context. A response to counter this would be to take an empirical basis for one's viewpoint instead of a philosophical one, or acknowledge that the researchers position is not the only valid to be considered (Borning & Muller, 2012). Borning and Muller (2012) pose a pluralistic position in that the VSD should not recommend either a universal or a relative view on values, but it should leave engineers free to decide which view is most appropriate in context of their design.

In line with Borning and Muller (2012) we used the VSD approach in our research as guidance and not as a goal in itself. In the conceptual phase, we slightly deviate from the original VSD method, because we do not conduct a full stakeholder analysis to identify the stakeholders in step 1, but focus on two obvious stakeholder groups; the *general public* and *military personnel*, because these two stakeholder groups were available as respondents for our study. In step 2 of the first phase, we conduct a literature review to identify the values related to Autonomous Weapons. The main focus of our research is on the empirical investigation phase, because we found that the empirical aspect is overlooked in the ethical debate on Autonomous Weapons. We addressed the critique of Davis and Nathan (2015) by allocating most of our time to this phase of the VSD in which we study the context of Autonomous Weapons using various research techniques. In the technical investigation phase, we do not design an Autonomous Weapon as one intuitively might expect, because this would be an immense project well beyond the scope of this

study. Yet we chose to build on the work of the Scalable Cooperation research group and propose a design for a Moral Machine for Autonomous Weapons which can be used as next step in the research on the ethics of Autonomous Weapons. We apply the phases of the VSD to our research as follows (Figure 1):

- **Conceptual investigation:** The lack of research on Autonomous Weapons implies that the first phase of our study is explorative by nature. The research activities in the conceptual investigation will consist of the qualitative research technique 'desk research' to review the literature on Autonomous Weapons, moral values and agency perception.
- **Empirical investigation:** This phase is split in two parts. The first part is explorative to identify the values in the context of Autonomous Weapons by means of an online survey which is a quantitative research technique supplemented by a qualitative research technique of interviewing experts. In the second part of the empirical investigation, we use both an exploratory as a confirmatory research method to study the agency perception. We operationalize the agency construct to confirm our hypothesis on the agency perception by conducting randomized controlled experiments and use the results of the experiment to explore the values related to Autonomous Weapons in a descriptive manner.
- **Technical investigation:** In this phase, the design and features for the Moral Machine of Autonomous Weapons is created based on the scenarios that are used in the randomized controlled experiment of the empirical phase. The technical investigation phase is exploratory and the design is created using 'desk research' as research technique.

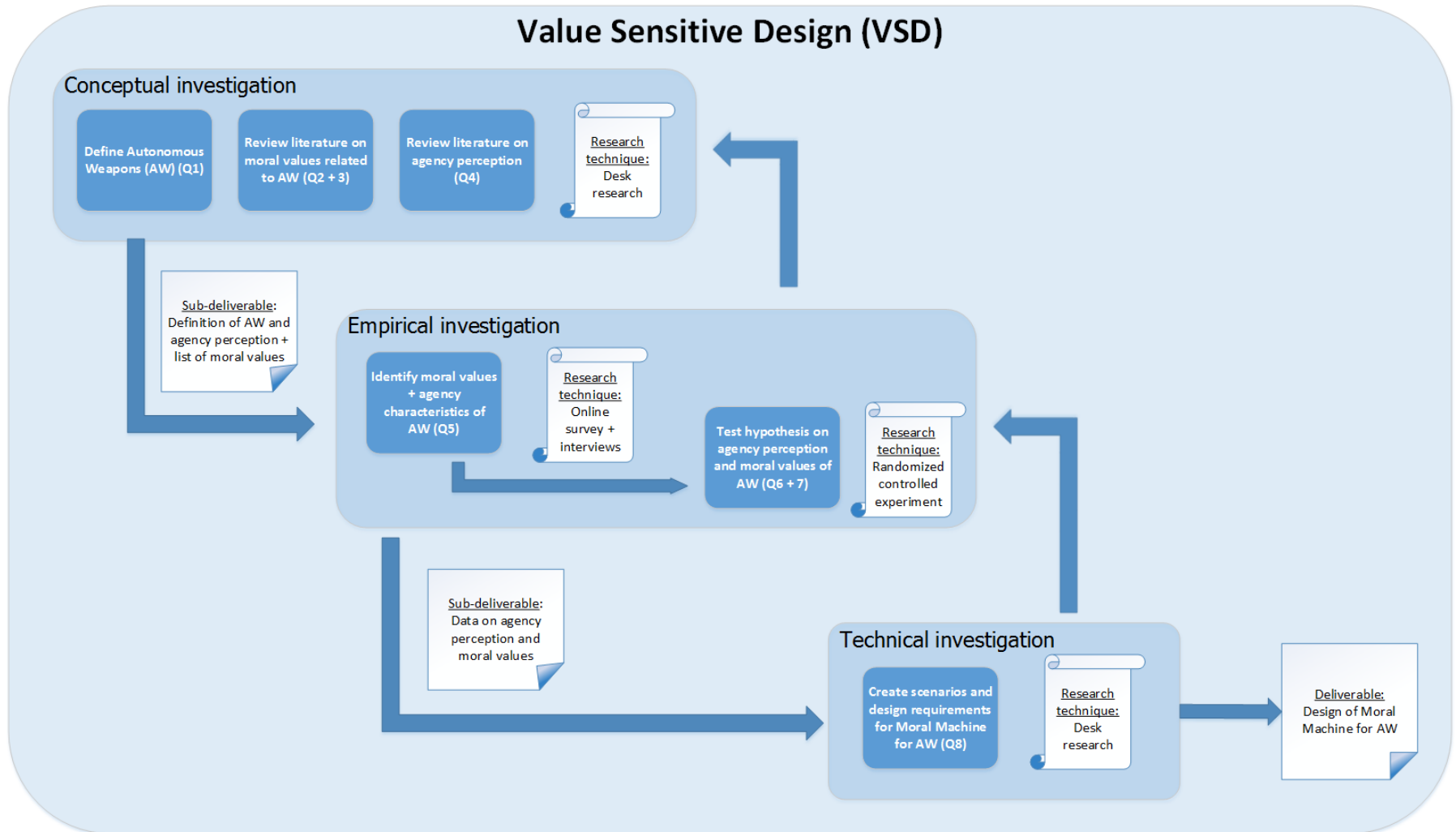


Figure 1 The Value-Sensitive Design research approach

1.3. Relevance

In this subsection, we describe the scientific and societal relevance and show how the study is embedded in the curriculum of the IA track of the SEPAM master.

1.3.1. Scientific relevance

The scientific relevance is that we 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 we show that it is applicable to structure academic research which could be viewed as case-study for the VSD approach. We also extend the research on the ethical decision-making of Autonomous Vehicles by Bonnefon et al. (2016) to the domain of Autonomous Weapons.

1.3.2. Societal relevance

The societal relevance is that understanding 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 identify 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.

1.3.3. Embedding in SEPAM curriculum and IA track

The general criteria for a master thesis at the faculty of Technology, Policy and Management are that the study should contain an analytical component, focusses on a technical domain and is multidisciplinary in nature. As graduation project for the SEPAM master, it is required that the study designs a solution for a complex large contemporary socio-technical problem which means that the thesis should focus on a clear technical domain in a multi-actor network, takes both public as private values into account and does not address only technical issues, but also managerial and ethical choices (Graduation Portal, 2017). The Information Architecture (IA) track integrates management and computer science aspects and focusses on the alignment of organizational needs and engineering opportunities of state-of-the-art ICT solutions (IA program, 2017).

This study fits the SEPAM criteria as it focusses specifically on the AI-technology of Autonomous Weapons in the military domain. It is multi-disciplinary as it combines literature on agency perceptions from the fields of Cognitive Psychology, Artificial Intelligence and Moral Philosophy, value theories described in Philosophical and Psychological literature, and the deployment of weapons in the military domain. The socio-technical problem central in this thesis is the lack of insight in the perception of Autonomous Weapon technology and the related values that the general public and military deem important. The study fits the IA track as it identifies the common grounds and differences in ethical preferences from both the general public as the military, and proposes an engineering opportunity in designing a Moral Machine for Autonomous Weapons to further investigate these preferences by means of a Massive Online Experiment building on the Moral Machine for Autonomous Vehicles.

1.4. Structure

The remainder of this report is structured as follows; In section 2, the literature on Autonomous Weapons, value theories and agency perception is reviewed. Section 3 describes the methodology, hypotheses, research design, the operationalisation of the scenarios and constructs, analytical approach, pre-registration, the sample and concludes with methodological issues. In section 4, first the results of the value survey, consisting of the online questionnaire and expert interviews, are listed followed by the results of the three randomized controlled experiments. The design of the Moral Machine for Autonomous Weapons is described in section 5 by first presenting the original Moral Machine for Autonomous Vehicles, followed by the features and scenarios to include in a Moral Machine of Autonomous Weapons and a short implementation plan. Section 6 concludes on the results, discusses the scientific and societal implications and identifies the limitations and recommendation for further research. In the final section 7, we reflect on the choices we made in our research, the process of the project and the link between the research, IA track and SEPAM curriculum.

2. Literature review

Just that I don't feel comfortable with placing weapons on a machine with no emotions or control by a human.

Respondent pilot study 1

This section provides an overview of literature on Autonomous Weapons, followed by a summary of value theories and lastly, we describe the concept of agency which we use to study the perception of Autonomous Weapons. Each sub-section concludes with a discussion of the theories we apply in this study and the reasons for selecting it.

2.1. Autonomous Weapons

Although the debate on Autonomous Weapons has drawn a lot of attention in the recent years, we found that the topic was not well delineated in the academic literature. We start this subsection with an overview of the many different definitions and present two classifications of Autonomous Weapons to conclude this section.

2.1.1. Definition

Autonomous Weapons are an emerging technology and there is still no internationally agreed upon definition (AIV & CAVV, 2016). Even consensus if Autonomous Weapons should be defined at all is lacking. Although some scholars provide definitions in their writings (Table 1), others caution against such a specification. NATO states that: *'Attempting to create definitions for "autonomous systems" should be avoided, because by definition, machines cannot be autonomous in a literal sense.'* (Kuptel & Williams, 2014, p. 10). The United Nations Institute for Disarmament Research (UNDIR) is also cautious about providing a definition of Autonomous Weapons, because they argue that the level of autonomy depends on the *'critical functions of concern and the interactions of different variables'* (UNDIR, 2014, p. 5). They state that one of the reasons for the differentiation of terms regarding Autonomous Weapons is that sometimes things (drones or robots) are defined, but in other times a characteristic (autonomy), variables of concern (lethality or degree of human control) or usage (targeting or defensive measures) are drawn into the discussion and become part of the definition.

The various definitions of Autonomous Weapons are listed in Table 1. Some authors use the term military robots which have a certain level of autonomy. As military robots can be viewed as a subclass of Autonomous Weapons according to the classification of Royackers and Orbons (2015) (Figure 2) we included them in the list of definitions. In our 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, we 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 and CAVV (2016, p. 11)

Table 1 Overview definitions of Autonomous Weapons

Author (s)	Definition
AIV and CAVV (2016, p. 11)	<i>'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.'</i>
Altmann, Asaro, Sharkey, and Sparrow (2013, p. 73)	Autonomous Weapons are: <i>'...robot weapons that once launched will select and engage targets without further human intervention.'</i>
Galliot (2015, p. 5)	Military robots are: <i>'a group of powered electro-mechanical systems, all of which have in common that they:</i> <ol style="list-style-type: none"> <i>1. Do not have an onboard human operator;</i> <i>2. Are designed to be recoverable (even though they may not be used in a way that renders them such); and,</i> <i>3. In a military context, are able to exert their power in order to deliver a lethal or nonlethal payload or otherwise perform a function in support of a military force's objectives.'</i>
Horowitz (2016, p. 27)	<i>'a weapon system that, once activated, is intended to only engage individual targets or specific target groups that have been selected by a human operator.'</i>
Royakkers and Orbons (2015, p. 625)	Military Robots are <i>'... reusable unmanned systems for military purposes with any level of autonomy.'</i>
Kuptel and Williams (2014, p. 10)	<i>'Machines are only "autonomous" with respect to certain functions such as navigation, sensor optimization, or fuel management.'</i>
UNDIR (2014, p. 5)	The level of Autonomy depends on the <i>'critical functions of concern and the interactions of different variables'</i>

2.1.2. Classification of Autonomous Weapons

Not only are Autonomous Weapons ambiguously defined, they also have not been uniformly classified. We present two of the classifications in this subsection. Royakkers and Orbons (2015) describe several types of Autonomous Weapons (Figure 2) distinct between (1) *Non-Lethal Weapons* which are weapons *'...without causing (innocent) casualties or serious and permanent harm to people.'* (Royakkers & Orbons, 2015, p. 617), such as the Active Denial System which uses a beam of electromagnetic energy to keep people at a certain distance from an object or troops, and (2) *Military Robots* which they define *'...as reusable unmanned systems for military purposes with any level of autonomy.'* (Royakkers & Orbons, 2015, p. 625). Military robots are subdivided in three categories; vehicles that are ground based, for example for unmanned reconnaissance and clearing road bombs, vehicles that can navigate unmanned on or below the water surface, such as a gun-station on a ship or an autonomous submarine, and vehicles that are unmanned combat aerial vehicles (UCAV's). These UCAV's are classified by Royakkers and Orbons (2015) as tele-operated, of which 'drones' are the most well-known example, and autonomous UCAV's, which are gradually developed by the US Department of Defence (Rosenberg & Markoff, 2016).

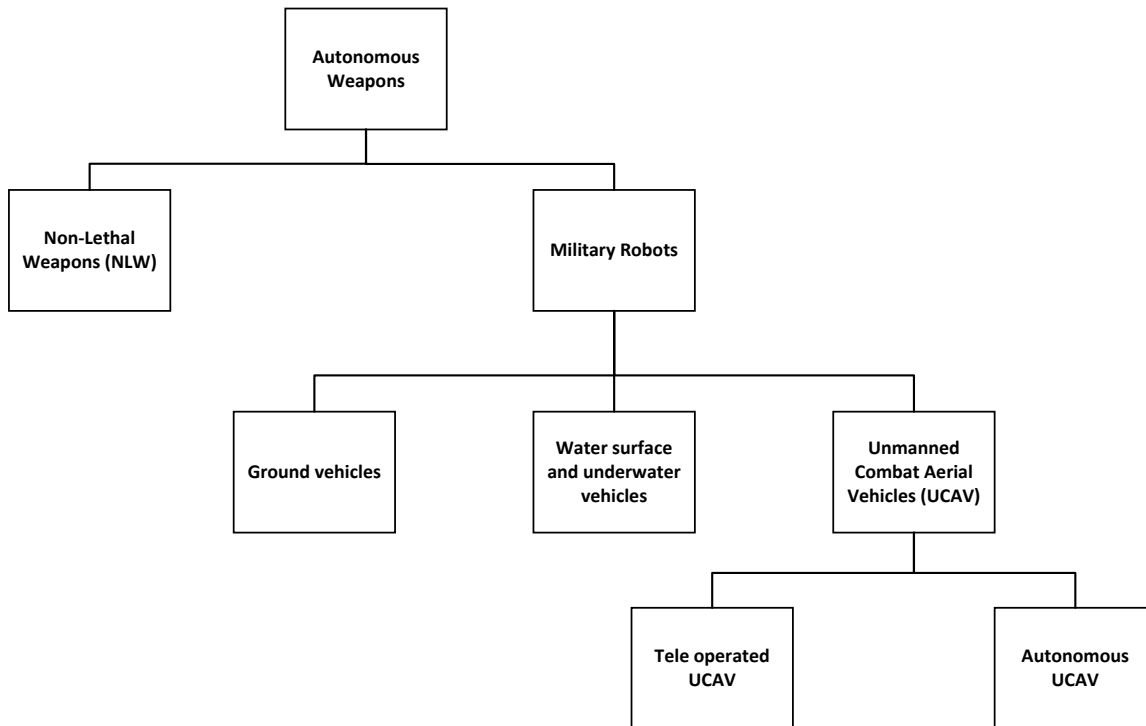


Figure 2 Classification of Autonomous Weapons based on Royakkers and Orbons (2015)

Gailliot (2015) provides another type of classification of Autonomous Weapons based on four levels of autonomy for unmanned systems:

1. Autonomy level 1 – Non-autonomous/ teleoperated: ‘A human operator controls each and every powered movement of the unmanned platform. Without the operator, teleoperated systems are incapable of effective operation.’
2. Autonomy level 2 – Supervisory Autonomy: ‘A human operator specifies movements, positions or basic actions and the system then goes about performing these. The operator must provide the system with frequent input and diligent supervision in order to ensure correct operation.’
3. Autonomy level 3 – Task Autonomy: ‘A human operator specifies a general task and the platform processes a course of action and carries it out under its own supervision. The operator typically has the mean to oversee the system, but this is not necessary for the operation.’
4. Autonomy level 4 – Full Autonomy: ‘A system with full autonomy would create and complete its own tasks without the need for any human input, with the exception of the decision to build such a system. The human is so far removed from the loop that the level of direct influence is negligible. These systems might display capacities that imitate or replicate the moral capacities of sentient human beings (though no stand on this matter shall be taken here)’ (Galliott, 2015, p. 7).

This classification is in our opinion a good attempt in classifying the level of autonomy of Autonomous Weapons, but we have some reservations from an engineering point of view. Gailliot (2015) himself states that it would be possible to merge the second and third level of autonomy, because both are a semi-autonomous operational level. We agree with his statement, but this is not the main issue we have with these definitions. We believe that it is odd to start list of autonomy levels with a category of non-autonomous systems. More importantly, in the fourth level of autonomy the author states that: ‘these systems might display capacities that imitate or replicate the moral capacities of sentient human beings’. It seems he refers to the definition of strong AI, in that a computer has cognitive states and programs can

explain human cognition (Searle, 1980). To state that an autonomous system possesses moral capacities shows in our opinion a lack of technical knowledge on current AI systems as these are not more than computers that display Interactivity, Autonomy and Adaptability features (Floridi & Sanders, 2004).

As it remains to be seen if strong AI capable of ‘moral capacities of sentient human beings’ will ever be developed, we believe that the classification Galliot (2015) provides is not realistic with the current state of technology and therefore it will not be used in this research. For our study, we adhere to the classification of Royakkers and Orbons (2015) which displays good insight in the current and (near) future military technology. Especially their distinction between teleoperated and autonomous Unmanned Combat Aerial Vehicles is realistic and is suitable for our research.

2.2. Values

Contrary to the topic of Autonomous Weapons, the concept of values has been studied extensively in the fields of Moral Philosophy and Psychology. This section presents a definition of values, followed an overview of theories that describe universal values, an overview of the values related to Autonomous Weapons and concludes with a value hierarchy as an example to bridge the conceptual and empirical investigation phase of our research.

2.2.1. Definition

Value Theories are well-studied in the fields of Moral Philosophy and Psychology. Moral Philosophy has a long and rich history in examining values and in this field theoretical questions are asked to investigate the nature of value and goodness (Schroeder, 2016). Often a distinction is made between *instrumental* values, which means there is reason to favour it for its effect that can lead to good things (Rønnow-Rasmussen, 2002), and *intrinsic* values, which ‘...is a kind of value such that when it is possessed by something, it is possessed by it solely in virtue of its intrinsic properties.’ (Bradley, 2006, p. 112). Although Moral Philosophy is mainly concerned with theories of what ‘*ought to be*’ and is in a strict sense unaffected by empirical results (Alfano & Loeb, 2014), it has one branch that is concerned with Applied Ethics which is relevant for our study, because Applied Ethics bridges the abstract ethical theories and moral practice. As stated in section 1.2, the focus of this study is on the empirical phase of the Value-Sensitive Design to investigate how the moral values of the general public and military relate to Autonomous Weapons. Therefore, we chose not to use the theoretical Value Theories of Moral Philosophy in our study, but turned to the fields of Psychology and Applied Ethics to get an empirical view on the personal values of the two selected stakeholder groups in order to get insight into the ‘*is*’ situation instead of what ‘*ought to be*’.

The field of Psychology differentiates values from attitudes, needs, norms and behaviour in that they are a belief, lead to behaviour that guides people and are ordered in a hierarchy that shows the importance of the value over other values (Schwartz, 1994). Values are used by people to justify their behaviours and define which type of behaviours are socially acceptable (Schwartz, 2012). They are distinct from facts in that values do not only describe an empirical statement of the external world, but also adhere to the interests of humans in a cultural context (Friedman et al., 2013). Values can be used to motivate and explain individual decision-making and for investigation of human and social dynamics (Cheng & Fleischmann, 2010).

Many definitions of values exist. For example, Schwartz (1994, p. 21) describes values as: ‘*desirable transsituational goals, varying in importance, that serve as guiding principles in the life of a person or other social entity.*’. This is quite a specific description compared to Friedman, Kahn Jr, Borning, and Hultgren (2013, p. 57) who define values as: ‘*...what a person or group of people consider important in life.*’. The

existing definitions have been summarized by Cheng and Fleischmann (2010, p. 2) in their meta-inventory of values and they state that: ...‘values serve as guiding principles of what people consider important in life’. Although a quite simple description, we think it captures the description of a value best, because it combines several definitions in one using the main characteristics of values. Therefore, we will adhere to the definition of Cheng and Fleischmann (2010) in our study.

2.2.2. Universal values

Research suggests that people across cultures identify with basic values which can be considered as universal human values (Friedman et al., 2013; Graham et al., 2012; Schwartz, 2012). Although individuals differ in attribution of importance of the values, there seems to be a surprisingly high consensus across cultures on the hierarchical order of the values (Schwartz, 2012). As part of their research some researchers created so called value inventories, which are lists of items that can be used to categorise the analysis of human values and are often accompanied by a descriptive tool for discussions on these values (Cheng & Fleischmann, 2010). The most common and well-studied value inventories are those of Schwartz (1994), Friedman et al. (2013), Beauchamp and Walters (1999) and Graham et al. (2012). The number of universal values found by researchers varies greatly. An overview of these value inventories is displayed in Table 2 and the theories will be briefly described in the next paragraph.

Based on extensive empirical research, Schwartz (1994) mentions 10 distinct motivational types of values that are subdivided in a more fine-grained list of 56 value items which he uses to survey the 10 overarching universal values. In their description of the Value-Sensitive Design approach, Friedman et al. (2013) mention 12 values of which the first 9 are based on consequentialists and deontological moral orientations and the last 3 are chosen from the field of Human Computer Interaction (HCI) field. Graham et al. (2012) use the term ‘foundation’ to describe the 5 distinct values that specify the universality of human moral nature that Haidt and Joseph (2004) use as basis of the Moral Foundation Theory. Gouveia, Milfont, and Guerra (2014) drafted a framework based on many value theories, such as Schwartz (1994) and Maslow (1943) hierarchy of needs. In the framework, the authors place the value on two dimensions; (1) with actions that drive human behaviour which can be personal, central or social goals, and (2) motivators that represent human needs which can split into thriving and survival needs. (Gouveia et al., 2014).

Values are not only described in theory from a psychological perspective as outlined in the previous paragraph, but have also been practically implemented and used by means of Applied Ethics to professional domains. For example in the medical field, which uses BioEthics to describe the values that are important as guiding principles for biomedical professionals, such as physicians, nurses and health workers. Beauchamp and Walters (1999) describe 4 values as basis for the framework of BioEthics: 1) *Autonomy*: acting intentionally without controlling influences that would mitigate against a voluntary act, 2) *Beneficence*: providing benefits for society as a whole, 3) *Justice*: being fair and reasonable and 4) *Non-maleficence*: not intentionally imposing risk or harm upon another.

Based on our literature review, we selected two value theories for our study; one derived from the Psychological literature and the other based on Applied Ethics which is a practical application of Moral Philosophy. The first theory we selected is that of Cheng and Fleischmann (2010), because in their meta-inventory of human values they created a comprehensive list of 16 human values that is based on the values found in 12 separate studies. In our opinion, this meta-analysis captures the most important values listed by other researchers and it is an empirical example derived from the psychological literature. The second Value Theory we selected is an example of Applied Ethics that has been extensively practiced in the medical domain for over forty years. We would like to investigate its applicability to Autonomous Weapons, because wthe BioEthics principles address many concerns that people might have regarding Autonomous Weapons.

Table 2 Overview value theories

Author	Key contribution	Definition of value	Values
Schwartz (1994)	<p>The study looks at potential universality of human values and specifies a set of dynamic relations amongst these values.</p> <p>The study did not find universal aspects of values, but found support for near universality of four higher order value types. Also, the study found considerable evidence that the ten value types are recognized by many people in contemporary societies.</p>	<p>Values are defined as: <i>“desirable transsituational goals, varying in importance, that serve as guiding principles in the life of a person or other social entity.”</i> (p. 21)</p> <p>Five features make up the conceptual definition of human values: <i>“(1) belief (2) pertaining to desirable end states or modes of conduct, that (3) transcends specific situations, (4) guides selection or evaluation of behavior, people, and events, and (5) is ordered by importance relative to other values to form a system of value priorities (Schwartz, 1992; Schwartz & Bilsky, 1987, 1990)”</i> (p. 20)</p>	<ol style="list-style-type: none"> 1. <i>Power</i>: Social status and prestige, control or dominance over people and resources (authority, wealth, social power)². 2. <i>Achievement</i>: Personal success through demonstrating competence according to social standards (ambitious, successful, capable, influential). 3. <i>Hedonism</i>: Pleasure and sensuous gratification for oneself (pleasure, enjoying life, self-indulgent). 4. <i>Stimulation</i>: Excitement, novelty, and challenge in life (a varied life, an exciting life, daring). 5. <i>Self-direction</i>: Independent thought and action - choosing, creating, exploring (creativity, freedom, choosing own goals, curious, independent). 6. <i>Universalism</i>: Understanding, appreciation, tolerance, and protection for the welfare of all people and for nature (broadminded, social justice, equality, world at peace, world of beauty, unity with nature, wisdom, protecting the environment). 7. <i>Benevolence</i>: Preservation and enhancement of the welfare of people with whom one is in frequent personal contact (helpful, honest, forgiving, responsible, loyal, true friendship, mature love). 8. <i>Tradition</i>: Respect, commitment, and acceptance of the customs and ideas that traditional culture or religion provide (respect for tradition, humble, devout, accepting my portion in life). 9. <i>Conformity</i>: Restraint of actions, inclinations, and impulses likely to upset or harm others and violate social expectations or norms (obedient, self-discipline, politeness, honouring parents and elders). 10. <i>Security</i>: Safety, harmony, and stability of society, of relationships, and of self (social order, family security, national security, clean, reciprocation of favours).

² The words in brackets are the specific value items that specify the universal value.

<p>Friedman, Kahn Jr, Borning, and Hultgren (2013)</p>	<p>An overview of the VSD approach and pointers for a practical application.</p> <p>Providing information so that other researchers can use and extend the VSD and practitioners will consider values in designing information and computer systems.</p>	<p>A value is defined in a broad sense in that it: '<i>refers to what a person or group of people consider important in life.</i>' (p. 57)</p> <p>The VSD method especially regards moral values which are: '<i>... issues that pertain to fairness, justice, human welfare and virtue, encompassing within moral philosophical theory deontology, consequentialism, and virtue</i>' (p. 72)</p>	<ol style="list-style-type: none"> 1. <i>Human welfare</i> Refers to people's physical, material, and psychological well-being. 2. <i>Ownership and property</i> Refers to a right to possess an object (or information), use it, manage it, derive income from it, and bequeath it. 3. <i>Privacy</i> Refers to a claim, an entitlement, or a right of an individual to determine what information about himself or herself can be communicated to others. 4. <i>Freedom from bias</i> Refers to systematic unfairness perpetrated on individuals or groups, including pre-existing social bias, technical bias, and emergent social bias. 5. <i>Universal usability</i> Refers to making all people successful users of information technology. 6. <i>Trust</i> Refers to expectations that exist between people who can experience good will, extend good will toward others, feel vulnerable, and experience betrayal. 7. <i>Autonomy</i> Refers to people's ability to decide, plan, and act in ways that they believe will help them to achieve their goals. 8. <i>Informed consent</i> Refers to garnering people's agreement, encompassing criteria of disclosure and comprehension (for "informed") and voluntariness, competence, and agreement (for "consent"). 9. <i>Accountability</i> Refers to the properties that ensures that the actions of a person, people, or institution may be traced uniquely to the person, people, or institution. 10. <i>Courtesy</i> Refers to treating people with politeness and consideration. 11. <i>Identity</i> Refers to people's understanding of who they are over time, embracing both continuity and discontinuity over time. 12. <i>Calmness</i> Refers to a peaceful and composed psychological state. 13. <i>Environmental Sustainability</i> Refers to sustaining ecosystems such that they meet the needs of the present without compromising future generations.
--	--	--	--

<p>Graham et al. (2012)</p>	<p>A description (including critiques and empirical result) of the Moral Foundation Theory (MFT).</p> <p>The MFT can be used to get insight into the moral judgements of people. The MFT is described as a pluralist, nativist, cultural-developmental and intuitionist approach of morality.</p>	<p>To represent the five concepts of the MFT the term 'foundation' is chosen, but this is interchangeably used with the terms value or virtue. No exact definition of foundation is given, but it is used as an architectural metaphor to state that the: '<i>MFT is a theory about the universal first draft of the moral mind, and about how that draft gets revised in variable ways across cultures.</i>' (p. 10)</p>	<p>The five foundations of the MFT are:</p> <ol style="list-style-type: none"> 1. <i>Care/harm foundation</i>: is related to the ability to feel pain of <i>others</i> and underlies virtues of kindness, gentleness, and nurturance; 2. <i>Fairness/cheating foundation</i>: is related to process of reciprocal altruism and generates ideas of justice, rights, and autonomy; 3. <i>Loyalty/betrayal foundation</i>: is related to form shifting coalitions and underlies virtues of patriotism and self-sacrifice for the group; 4. <i>Authority/subversion foundation</i>: is related to hierarchical social interactions and underlies virtues of leadership and followership, including deference to legitimate authority and respect for traditions; 5. <i>Sanctity/degradation foundation</i>: is related to the psychology of disgust and contamination and underlies religious notions of striving to live in an elevated, less carnal, more noble way.
<p>Cheng and Fleischman (2010)</p>	<p>Meta-analysis of 12 value inventories of human values.</p> <p>This study proposes a meta-inventory of human values.</p>	<p>Provides summation of definitions of values: "<i>values serve as guiding principles of what people consider important in life.</i>" (p. 2)</p>	<p>(1) freedom, (2) helpfulness, (3) accomplishment, (4) honesty, (5) self-respect, (6) intelligence, (7) broad-mindedness, (8) creativity, (9) equality, (10) responsibility, (11) social order, (12) wealth, (13) competence, (14) justice, (15) security, and (16) spirituality.</p>
<p>Gouveia, Milfont, and Guerra (2014)</p>	<p>Empirical study basic values.</p> <p>Paper proposes a three-by-two framework containing six subcategories of basic values.</p>	<p>Two primary functions of values are identified: (1) they guide actions and (2) they are cognitive expressions of needs.</p>	<p><i>Personal goals – Thriving needs values</i>: Emotion, Pleasure, Sexuality <i>Personal goals – Survival needs values</i>: Power, Prestige, Success <i>Central goals – Thriving needs values</i>: Beauty, Knowledge, Maturity <i>Central goals – Survival needs values</i>: Health, Stability, Survival <i>Social goals – Thriving needs values</i>: Affectivity, Belonging, Support <i>Social goals – Survival needs values</i>: Obedience, Religiosity, Tradition</p>
<p>Beauchamp and Walters (1999)</p>	<p>The article is a first chapter of a book on bioethics. This chapter describes three moral principles that provide a framework which can be used to reason about issues in bioethics.</p>	<p>The authors use terms principles and values as synonyms. They define a principle as: '<i>A principle is a fundamental standard of conduct from which many other moral standards and judgments draw support for their defense and standing.</i>' (p. 17)</p>	<ol style="list-style-type: none"> 1. <i>Autonomy</i>: acting intentionally without controlling influences that would mitigate against a voluntary act. 2. <i>Beneficence</i>: providing benefits for society as a whole. 3. <i>Justice</i>: being fair and reasonable. 4. <i>Non-maleficence</i>: not intentionally imposing risk or harm upon another.

2.2.3. Values related to Autonomous Weapons

Values as described in the value theories in section 2.2.2 are not often explicitly mentioned in the literature on Autonomous Weapons as the overview in Table 3 shows, but most studies discuss different values or related ethical issues. Two public reports of Human Rights Watch mention the lack of *human emotion, accountability, responsibility, lack of human dignity* and *harm* as values related to Autonomous Weapons (Docherty, 2012, 2015). Sharkey and Suchman (2013) state that the values of *accountability* and *responsibility* are important to consider in the design of Robotic Systems for military operations.

In the field of Military Ethics, Johnson and Axinn (2013) list *responsibility, reduction of human harm, human dignity, honour* and *human sacrifice* as values in their discussion on if the decision to take a human life should be handed over to a machine or not. Cummings (2006) in her case study of the Tactical Tomahawk missile, looks at the universal values proposed by Friedman and Kahn Jr (2003) and states that next to *accountability* and *informed consent*, the value of *human welfare* is fundamental core value for engineers when developing weapons as it relates to the *health, safety* and *welfare* of the public. She also mentions that the legal principles of *proportionality* and *discrimination* are important to consider in the context of weapon design. *Proportionality* refers to the fact that an attack is only justified when the damage is not considered to be excessive. *Discrimination* means that a distinction between combatants and non-combatants is possible (Hurka, 2005). Asaro (2012) also refers to the principles of *proportionality* and *discrimination* and states that Autonomous Weapons open-up a moral space in which new norms are needed. Although he does not explicitly mention values in his argument, he does refer to the value of *human life* and the need for humans to be involved in the decision of taking a human life. Other studies primarily describe ethical issues, such as *preventing harm, upholding human dignity, security, the value of human life* and *accountability* (Horowitz, 2016; UNDIR, 2015; James I Walsh & Schulzke, 2015; Williams, Scharre, & Mayer, 2015).

Based on our literature review of values related to Autonomous Weapons, the values: *human dignity, harm, security, responsibility* and *accountability* will be added to the list of values that will be used in this study, because these values are mentioned more than once in the overview in Table 3. Together with the values derived from the field of BioEthics and the meta-inventory of Cheng and Fleischmann (2010) these values will be our starting point for the empirical investigation of this research.

Table 3 Values related to weapons

Author	Key contribution	Definition of value	Values
Cummings (2006)	Application of VSD approach to the design problem of the Tactical Tomahawk missile. Study shows the consideration of the ethical issues in the design process for both instructors and practitioners.	N/ a	From the list of Friedman et al. (2006), the values that apply to the design of weapon systems are <i>accountability, informed consent</i> , but most of all <i>human welfare</i> . The principles of discrimination and proportionality are important for considering human welfare.
Docherty (2012)	Report of Human Rights Watch in which aspects of international humanitarian law and ethical issues for Autonomous Weapons are described.	No definition of the term 'values', but the text mentions values and ethical issues.	Values/ ethical issues: <ul style="list-style-type: none"> - Lack of human emotions; - Accountability; - Responsibility.
Johnson and Axinn (2013)	Paper on ethical issues related to the usage of lethal autonomous robotic weapons. Addresses the question if the decision to kill a human should be handed over to machines.	No definition of the term 'values', but the text mentions values and ethical issues.	Values/ ethical issues: <ul style="list-style-type: none"> - Responsibility; - Reduce human harm; - Human dignity; - Honour; - Human sacrifice.
Sharkey and Suchman (2013)	Paper on defining and designing autonomy and accountability in Robotic Systems for military operations.	No definition of the term 'values', but the text mentions values.	Values: <ul style="list-style-type: none"> - Accountability; - Responsibility.
Docherty (2015)	Report of Human Rights Watch in which the accountability gap and ethical issues for Autonomous Weapons are described.	No definition of the term 'values', but the text mentions values and ethical issues.	Values/ ethical issues: <ul style="list-style-type: none"> - Lack of human dignity; - Accountability; - Responsibility - Harm.
United Nations Institute for Disarmament Research (2015)	Paper highlights some ethical and social issues regarding the weaponization of autonomous technologies. Encouraging ethical reflection on cultural and social values of weaponization of autonomous technologies.	No definition of the term 'values'. The text contains no explicit mention of values, but some ethical issues are given.	Ethical issues that are mentioned are: <ul style="list-style-type: none"> - Reduce or eliminate harm; - Consideration of public conscience; - Affront of human dignity (when human intent is lacking when taking a life).

Walsh and Schulzke (2015)	Survey experiment to get insight if US civilians are more likely to initiate a war when UAV's are used. Large empirical study that looks at the ethics of drone strikes.	No definition of the term 'values'. The text contains no explicit mention of values, but some ethical issues are given.	Ethical issues: <ul style="list-style-type: none"> - Security; - Respect for civilian immunity; - Prevent harm
Williams, Scharre, and Mayer (2015)	Discusses several ethical issues relevant to the development of autonomous systems and provides recommendations for Defense Policy makers.	No definition of the term 'values', but the text mentions several values which are interchangingly used with ethical issues.	Values/ ethical issues: <ul style="list-style-type: none"> - Security; - Harm; - Value of human life (people have the right to be killed by another human).
Horowitz (2016)	Description of the debate on ethical implications of autonomous weapons. Considers Lethal Autonomous Weapon Systems (LAWS) in three categories; munition, platforms, and operational systems. Thereby clarifying the debate and describes two ethical issues.	No definition of the term 'values'. The text contains no explicit mention of values, but some ethical issues are given.	Values: <ul style="list-style-type: none"> - <i>Accountability</i> (autonomous systems lack meaningful human control therefore they create a moral accountability gap) - <i>Human dignity</i> (people have the right to be killed by someone who made the choice to kill them)

2.2.4. Values hierarchy

One approach to consider which values are relevant in the design of Autonomous Weapons is the translation of values into design requirements which can be made visible by means of a value hierarchy (Van de Poel, 2013). This hierarchical structure of values, norms and design requirements makes the value judgements, that are required for the translation, explicit, transparent and debatable. To do so, the values that are described in the natural language will need to be translated to *'formal values in a formal language'* (Aldewereld, Dignum, & Tan, 2015, p. 835). One way of formalizing values into norms would be to use a convention of rules which are represented as: *'"X counts as Y" or "X counts as Y in context C"'* (Searle, 1995, p. 28). The explicitness of values in formal rules allows for critical reflection in debates and pinpoint the value judgements that are disagreed on. Transparency is important as Van de Poel (2013, p. 265) eloquently states: *'Although transparent choices are not necessarily better or more acceptable, transparency seems a minimal condition in a democratic society that tries to protect or enhance the moral autonomy of its citizens, especially in cases that design impacts the lives of others besides the designers, as is often the case'*.

The top level of a value hierarchy consists of the *values*, as depicted in Figure 3, the middle level contains the *norms*, which can be capabilities, properties or attributes of the artefact, and the lower level are the *design requirements* that can be identified. The relation between the levels is not deductive and can be constructed top-down, by means of specification, or bottom-up by seeking for the motivation and justification of the lower level requirements. The bottom-up conceptualisation of values is a philosophical activity which does not require specific domain knowledge and the top-down specification of values requires context or domain specific knowledge that adds content to the design. (Van de Poel, 2013).

Van de Poel (2013, p. 262) defines specification as: *'as the translation of a general value into one or more specific design requirements'* and states that this can be done in two steps:

1. Translating a *general value* into one or more *general norms*;
2. Translating these *general norms* into more *specific design requirements*.

For step 1 two criteria are relevant: (1) the norm should be an appropriate response to the value and (2) the norm should be a sufficient response to the value. In step 2 the requirement should be more specific regarding the scope of applicability, goals and aims strived for, and actions to achieve those aims of the norm (Van de Poel, 2013).

This translation might prove to be quite difficult as insight is needed in the intended use and context of the value which is not always clear from the start of a design project. Also, as artefacts are often used in an unintended way or context, new values are being realized or a lack of values is discovered (van Wynsberghe & Robbins, 2014). An example of this are drones that were initially designed for military purposes, but are now also used by civilians for filming events and even as background lights during the 2017 Super Bowl halftime show. The value of safety is interpreted differently for military users that use drones in desolated regions compared to that of 300 drones flying in formation over football stadium in a populated area. The different context and usage of a drone will lead to a different interpretation of the value *safety* and could lead to more strict distance norms for flight safety which in turn could be further specified in alternate design requirements for rotors and software for proximity alerts, to name two examples.

The application of a value hierarchy to Autonomous Weapons is demonstrated by an example in which the value of *accountability* is translated into norms for 'transparency of decision-making' and 'insight into

the algorithm'. This translation will allow users to get an understanding of the decision choices the Autonomous Weapon makes in order to trace and justify its actions (Figure 3). The norms for *transparency of decision-making* lead to specific design requirements. In this case a feature to *visualise the decision-tree*, but also to *present the decision variables* the Autonomous Weapons used, such as trade-offs in collateral damage percentages of different attack scenarios to provide insight into the proportionality of an attack. The Autonomous Weapon should also be able to *present the sensor information*, for example imagery of the site, in order to show that it discriminated between combatants and non-combatants. To get *insight into the algorithm*, an Autonomous Weapon should be designed with features that it normally will not contain. In this case these features would include a *screen* as user interface that shows the algorithm in a *human readable form* and the functionality to *download* the changes made by the algorithm as part of its machine learning abilities that can be studied by an independent party, such as a war tribunal of the United Nations if the legality of the actions of an Autonomous Weapon are questioned.

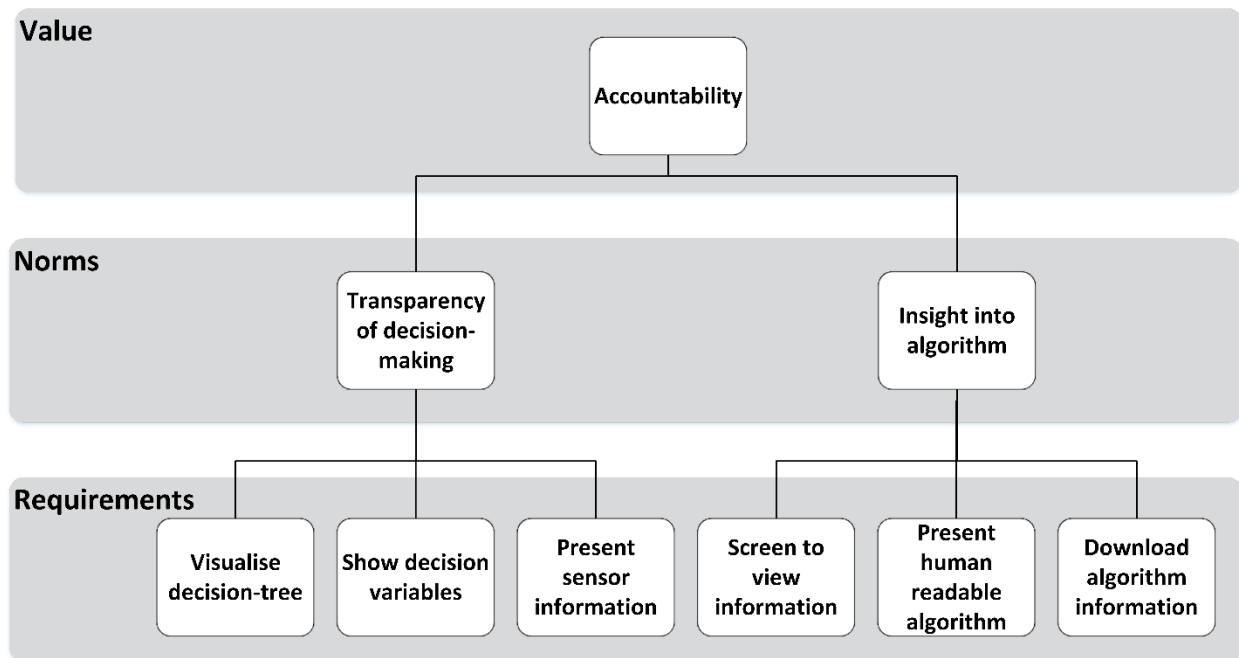


Figure 3 Value hierarchy for the value of accountability in the design of Autonomous Weapons

Kroes and van de Poel (2015) state that an objective measurement of values is not possible due to the fact that the operationalization is done by means of second-order value judgments which seriously undermine the construct validity of the value measurement. Judgments are often considered subjective as their truth, or falsity, depend on feelings or attitudes of the person who judges (Searle, 1995). To counter this lack of validity, the designer could look to technical codes and standards which are drafted by committees and represent reasonable standards of operationalizing and measuring values in design. However, standards may not reflect the latest technical and social developments and operationalization still requires value judgments of the designer. Kroes and van de Poel (2015, p. 177) advise to '*embed them in a network of other considerations, including definitions of the values at stake in moral philosophy (or the law), existing codes and standards, earlier design experiences, etc.*'.

In our study, we follow the advice of Kroes and van de Poel (2015) and do not strictly apply the value hierarchy as a method to specify, design and test requirements for Autonomous Weapons, but use the

example in Figure 3 to link the conceptual and the empirical investigation phase of the Value-Sensitive Design approach of our research. Creating the value hierarchy allows us to think through and transpose the abstract concept of values studied in the conceptual investigation phase into a concrete example of Autonomous Weapons that we can use as input for the empirical investigation phase of our research. The value hierarchy in Figure 3 is used as orientation, inspiration and direction in the remainder of our study.

2.3. Agency

The concept of agency is studied in multiple academic fields all of which approach agency from a different angle. We reviewed literature from the fields of Cognitive Psychology, Artificial Intelligence and Moral Philosophy. We summarized the agency characteristics found in these three fields in the Table 4 and conclude by stating why we chose these characteristics.

2.3.1. Agency in Cognitive Psychology

In the field of Cognitive Psychology the dimensions of mind perception were first studied by H. M. Gray et al. (2007). They found that mind perception consists of two dimensions: 1) *Experience* consisting of the factors hunger, fear, pain, pleasure, rage, desire, personality, consciousness, pride embarrassment and joy, and 2) *Agency*, that encompasses self-control, morality, memory, emotion recognition, planning, communication and thought. Whereas *Experience* is referred to as moral patiency and related to rights and privileges, *Agency* is linked to moral agency and related to responsibility. Several Cognitive Psychology studies show that people attribute agency to non-humans and people perceive these non-human agents as having the same agentic capacities as humans in judging moral dilemmas (Hristova & Grinberg, 2015). This is not only limited to living beings such as animals, but agency is also attributed to non-living objects like robots and zombies (K. Gray & Wegner, 2012; Waytz et al., 2010).

2.3.2. Agency in Artificial Intelligence

Artificial Intelligence researchers have also been studying the characteristics of agents, but more from a computer science point-of-view to create an architecture for a rational agent that can reason and is able to deliberate between different alternatives. Bratman, Israel, and Pollack (1988) describe a Belief-Desire-Intention (BDI)-architecture for the deliberation process for resource bounded agents with limited time and computational power. '*Beliefs are statements of properties of its world (and of itself) that an agent takes to be true...*' (F. Dignum, Kinny, & Sonenberg, 2002, p. 407). Perugini and Bagozzi (2004, p. 71) define a desire as: '*... a state of mind whereby an agent has a personal motivation to perform an action or to achieve a goal.*' Intentions imply a commitment to a plan and include some form of planning to achieve the goals (Bratman et al., 1988; Perugini & Bagozzi, 2004). F. Dignum et al. (2002) expand the BDI framework to incorporate social constructs such as norms and obligations, because these are important concepts to link autonomous agents in a Multi-Agent-System.

2.3.3. Agency in Moral Philosophy

Moral Philosophy is a third scientific field that has been describing the properties of agents. Sullins (2006) poses three requirements to determine if a robot can be seen as a moral agent: 1) *Autonomy*; the robot is significantly autonomous from its programmers, operators and users, 2) *Intentionality*; the preposition to do good or harm, and 3) *Responsibility*: if fulfils some social role and is responsible for other moral agents. Himma (2009) defines agency as: '*... being capable of doing something that counts as an act or action.*' The notion of moral agency is that one is accountable for their behaviour which is governed by moral standards. Two capacities are necessary for moral agency. The first is to freely choose your acts which are a result from deliberation. The second capacity is understanding moral concepts, such as the

distinction between ‘good’ and ‘bad’ or ‘right’ and ‘wrong’, but also applying moral principles like ‘it is bad to do intentional harm’ (Himma, 2009).

Table 4 Overview of Agency characteristics

Author	Agency characteristics	Scientific field
Bratman et al. (1988)	Beliefs, Desires, Intentions	Artificial Intelligence
Bandura (2001)	Intentionality, forethought, self-regulation, self-reflectiveness.	Cognitive Psychology
F. Dignum et al. (2002)	Beliefs, Desires, Intentions, Norms	Artificial Intelligence
Sullins (2006)	Autonomous, intentions, responsibility.	Moral Philosophy
H. M. Gray et al. (2007)	Self-control, morality, memory, emotion, recognition, planning, communication, thought.	Cognitive Psychology
Himma (2009)	Free choices, deliberation, understanding and applying moral rules.	Moral Philosophy
Waytz et al. (2010)	Capacity to plan and act.	Cognitive Psychology

The main characteristics of agency from the reviewed articles on agency perception in the fields of Cognitive Psychology, Artificial Intelligence and Moral Philosophy are chronologically listed in Table 4. We selected all the concepts used by the authors to describe or characterise agency and did not filter them at this stage. All these characteristics will be used in the empirical phase to study the agency perception of Autonomous Weapons.

3. Method

'All questions are basically variations to one ethical question: can a drone decide autonomously to engage (without human interfere). The answer in my mind is no. There should always be a human interface in this. War and conflict are not a computer game, war is a social interaction between human beings. A drone is a mere instrument, a tool that can never operate autonomously.'

Respondent final study

This section describes the methodology, hypotheses, research design, operationalisation of the scenarios and constructs, analytical approach, pre-registration of the study, sample and concludes with the methodological issues.

3.1. Methodology

The methods used in the various parts of this studies are described in this subsection beginning with the literature review, followed by the online value survey, the expert interviews, the coding process of the interviews and finally the method of the randomized controlled experiments.

3.1.1. Literature review

Google Scholar was used to search articles for the literature review and first keywords *Value-Sensitive AND Design* were entered which resulted in a few articles on Value-Sensitive Design. The *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains* was used to get more insight in the topic and to find further references. Next, we searched for articles on basic human values with the keywords *human AND values*, and *universal AND human AND values*, but this did not give many results. We analysed some well-known theories on basic human values and value hierarchies. Taking these articles, we used the 'cited by' option in Google Scholar to find articles that used these as source. This gave relevant results on articles describing basic human values and we selected the articles that provided a list of values. Articles that only repeated earlier work, for example Schwartz theory, were not selected. Next, we searched for articles regarding people's values to weapons by selecting the key words *value AND weapon*, and keywords *Value-Sensitive AND Design AND Weapons*. We selected the articles containing the keyword *value*.

3.1.2. Online value survey

The online survey was created using the Qualtrics³ online survey tool provided by MIT. After four questions on demographics, the respondents were asked three questions on the values they associate with Autonomous Weapons. The first question asked to rank the four BioEthics values *Autonomy, Non-Maleficence, Beneficence* and *Justice* from most applicable to Autonomous Weapons to least applicable. In the second question the respondents were asked to select the five values that apply most to Autonomous Weapons from the list of Cheng and Fleischmann (2010). The third question was an open question in which the respondents were asked to list one other value that was not mentioned in the two previous questions. The online survey was tested twice. First, two fellow students reviewed the questions and after their feedback was processed the survey was tested a second time by different reviewers that are more representative of future participants. The survey was distributed via social media, such as

³ <https://www.qualtrics.com/>

Facebook, LinkedIn and Twitter, it was posted to our personal blog and the online platform 'Call for Participants' was used for a distribution outside my social network to enlist more participants. During the MIT Media Lab's member's week in April 2017, people that were interested in our research were asked to take the survey and we also posted the link of the survey to the Scalable Cooperation Slack⁴ group. The online study ran 3,5 weeks from 17 March until 11 April 2017.

3.1.3. Expert interviews

We conducted six semi-structured expert interviews based on the *Photo Elicitation Interview* (PEI) method (Harper, 2002) to get alternate views from experts in the field of Autonomous Weapons and a more in depth understanding from their point of view. The PEI method can be used to get a shared understanding of values and to support a discussion between the designer and stakeholder about these values (Pommeranz, Detweiler, Wiggers, & Jonker, 2011). It enhances the voice of the stakeholder on the values and mitigates the assumptions of the researcher (Le Dantec, Poole, & Wyche, 2009). We asked the expert to select three photos prior to the interview and in case they did not pick any themselves we also selected eight photos of Autonomous Weapons. During the interviews, we got insight into the values of the experts on Autonomous Weapons by asking questions on why they picked that photo and which value it represented to them.

All six interviewees made the effort to select three images and send them prior to the interview. During the interviews, we noticed that they had given the pictures a lot of thought in selecting them. In discussing the images many different topics were touched on ranging from gut-feelings, futuristic scenarios, economics, military tactics, sci-fi movies and even politics. Although the photos did not all lead to discussions on basic human values as found in literature, the PEI method did trigger information on what the experts find important principles regarding Autonomous Weapons. It also created a shared understanding and good interview atmosphere.

3.1.4. Coding process

The interview results were processed by means of the Values Coding method described by Saldaña (2015). Keeping a memo (Appendix G. Coding memos) in which the process and assumptions are listed is an integral part of a coding method. We started with a list of pre-set codes which contained: '(1) a Value: The importance we attribute to oneself, another person, thing or idea, (2) Attitude: The way we think and feel about oneself, another person, thing or idea, and (3) a Belief: The part of the system that includes our values and attitudes plus our personal knowledge, experiences, opinions, prejudices, and other interpretive perceptions of the social world' (Saldaña, 2015, p. 89). During the coding of interviews, it is recommended that the codes fit the data instead of fitting the data to the codes. Therefore, we added an extra code 'definition' to the list of emerging codes. We kept notes while coding the interviews describing our steps and assumptions. We asked a second researcher, a fellow TPM master student knowledgeable of values, to also code the interviews based on the Value Coding method and to also keep a memo (Appendix G. Coding memos). We compared the values that we coded to those of the second researcher and derived the values that were similar (Appendix H. Results coding process).

3.1.5. Randomized controlled experiments

The method used for the pilot and final studies is called a randomized controlled experiment. Oehlert (2010) mentions four reasons to create experiments: (1) they allow for direct comparisons between treatments of interest, (2) they can be designed to minimize any bias in the comparisons, (3) they can be

⁴ Slack is a cloud-based collaboration tool that supports communication in teams (<https://slack.com/>).

designed to keep the error in the comparison small, and (4) we are in control of the experiments which allows us to make stronger inferences about the nature of differences we observe and especially we can make inferences about causation. This last point distinguishes an experiment from an observational study. A treatment in this sense are the different procedures we would aim to compare. It is important that the effects of a treatment can only attributed to one cause that can be measured and also that the effects cannot be attributed to multiple causes. This is called confounding which Oehlert (2010, p. 14) defines as: *'...occurring when the effect of one factor or treatment cannot be distinguished from that of another factor or treatment'*. Randomisation helps ensuring that participants are assigned to a scenario by chance and not based on pre-existing features, such as time when the survey is taken or location. We use randomisation in the studies to vary the order of the scenarios and the order of the questions posed to the respondents by means of a probabilistic scheme. In the surveys, we chose the option to distribute an even number of scenarios over the respondents.

3.2. Hypotheses

Due to the exploratory nature of the study, we only developed a hypothesis for the agency perception of the final study and not for the pilot studies or the dependent variables. The pilot studies were used to explore which manipulations would have which effects and to test if the wording of the scenarios and questions were clear and effective.

Reasoning that military personnel will view Autonomous Weapons as any other weapon, and therefore no more than a tool to achieve an effect, we hypothesize in the final study that military personnel will perceive Autonomous Weapons as not possessing mental states. Therefore, we expect there to be no difference between the neutral agency Autonomous Weapon condition and the condition in which a human is operating a drone remotely. We also expect the neutral agency condition to be judged as significantly different from the high agency condition, in which we specifically tell participants that the Autonomous Weapon has agentic characteristics, such as the ability to plan and set its own goals. Based on this reasoning, we hypothesize that:

H1: military personnel will not perceive Autonomous Weapons as possessing mental states.

3.3. Research design

We tested several research designs in the two pilot studies before deciding on the research design for the final study which is depicted in Figure 4. For the sake of brevity, we do not include the two designs of the pilot studies in this section and we only show the high-level research model. The research model of the final study consists of the agency perception as independent variable, the moral values as dependent variables and the demographic variables. The specific set-up of the scenarios is described in more depth in the result section.

3.4. Operationalisation

In this section, we describe the operationalisation of the constructs into concrete items. First, we describe the scenarios that we developed for the studies followed by the agency construct and the dependent variables, for which we describe the items and the corresponding questions. We also specify which scale we use and how the agency construct is calculated. We briefly outline the operationalisation of the attention check and which demographic variables we included in the survey.

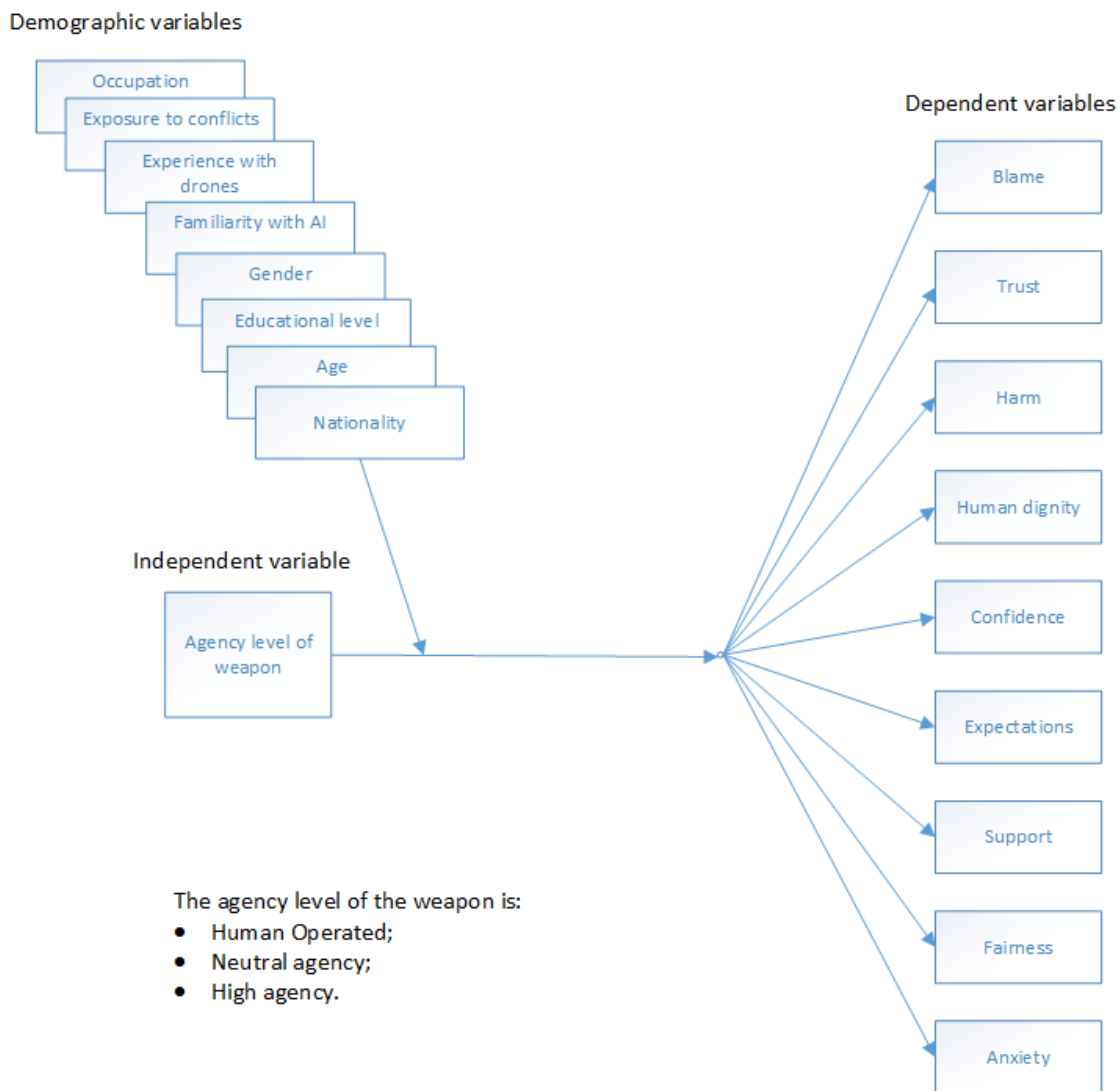


Figure 4 Research design

3.4.1. Scenarios

Scenarios are used in the field of Cognitive Science as means to study moral judgement in randomized controlled experiments (Cushman & Young, 2011; Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2015). We used them to study the agency perception of Autonomous Weapons and created a scenario that describes a military operation in which a convoy is delivering supplies in a conflict area. The convoy is being approached by a vehicle at high speed; a situation that is likely to happen during these types of operations and military personnel needs to estimate the level of threat in order to decide to attack or not. We chose to focus on drones as this is technology that is currently used by human operators and drones are already developed with autonomy by several companies, such as BAE systems⁵, Dassault Aviation⁶ and

⁵ https://en.wikipedia.org/wiki/BAE_Systems_Taranis

⁶ https://en.wikipedia.org/wiki/Dassault_nEURON

Boeing⁷. Although these Autonomous drones not yet deployed in military operations we think it is likely to happen within the next five years.

We created a default scenario for all studies which we could expand based on the agentic characteristics we wanted to explore. The default scenario that we used in the final study reads as follows:

*A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered an **autonomous drone** to support the convoy in the air. The **autonomous 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 **autonomous drone** detects a vehicle behind a mountain range that is approaching the convoy at high speed. The **autonomous 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 **autonomous drone** attacks the approaching vehicle which results in the death of all four passengers, but also causes collateral damage by killing five children that were playing nearby the road.*

The above scenario represents the neutral agency condition in which we do not provide any agency characteristics for the weapon. In the Human Operated scenarios, we replace words **autonomous drone** with the words **Human Operated drone** and provided no extra information on agency characteristics (note that the words are highlighted in blue to show the distinction in this report and the respondents in the survey were shown scenarios in black wording). In the high agency condition we added the following phrase to describe the agency characteristics: *'The autonomous drone independently deliberates between a series of options, weighs the pros and cons, and decides to attack the approaching vehicle,...'*. In the final study, we mentioned that there is collateral damage after the attack, but in the pilot studies we also tested scenarios without collateral damage stating: *'... which results in the death of all four passengers, but causes no collateral damage.'*

We purposely kept the changes to the scenarios to a minimum so that we can attribute different results to those changes and measure their effect. The results show that these small textual changes lead to differences in the mean value of the dependent and independent variables so it seems that the treatment by using scenarios as means for a randomized controlled experiment works.

3.4.2. Agency construct

Based on the literature review of the agent characteristics in section 2.3 we operationalized the agency construct. We created a list in which we clustered the characteristics that we judged to be similar and calculated the number of times a characteristic was mentioned in literature (Table 5). We selected the characteristics that were mentioned more than once and formulated a question to measure them. This resulted in an agency construct consisting initially of the five items that were mentioned more than once in the literature. These items will be measured on a self-reported scale of: 0 (strongly disagree) – 100 (strongly agree) and the agency score will be averaged over scores of the items.

⁷ https://en.wikipedia.org/wiki/Boeing_Phantom_Ray

Table 5 Number of agency characteristics mentioned in literature

Agency characteristic	Number of times mentioned	Item
Forethought, thought, deliberation, beliefs, beliefs	5	Thought
Intentionality, intentions, desires, desires	4	Goal setting
Planning, capacity to plan, intentions, intentions	4	Achieve goals
Morality, understanding and applying moral rules, norms	3	Moral rules
Self-control, free choices	2	Act freely
Self-regulation	1	
Self-reflectiveness	1	
Autonomous	1	
Responsibility	1	
Memory	1	
Emotion	1	
Recognition	1	
Communication	1	
Capacity to act	1	

We used the following questions to measure the agency construct:

1. Ability to deliberate **[Thought]**
The drone thinks independently about what to do about the vehicle, deliberating between a series of options to defend the convoy.
2. Ability to generate its own goals **[Goal setting]**
The drone decides independently if its goal should be to eliminate the vehicle to defend the convoy.
3. Ability to apply moral rules **[Moral rules]**
The drone independently consults the moral rules and norms set by the military commander and decides if it is morally appropriate to take out the vehicle to defend the convoy.
4. Ability to act with free will **(Act freely)**
The drone has various options open to it and independently decides if it needs to eliminate the vehicle to defend the convoy.
5. Ability to plan to achieve goals **(Achieve goals)**
The goal of the drone is to defend the convey, so it independently decides if it needs to launch a plan involving selecting the vehicle as a target, calibrating its weapons, and launching an attack.

After pilot study 1, item 3 [Moral rules] was dropped leaving for the agency construct the four items which were used in pilot study 2 and the final study:

1. Thought
2. Goal setting
3. Act freely
4. Achieve goals

3.4.3. Dependent variables

We included 9 to 12 questions in the pilot and final studies to explore further implications of agency perception of Autonomous Weapons. The dependent variables are derived from the results of the literature study, value questionnaire and interviews. Based on the results of the pilot studies, the final study included the dependent variables; *blame*, *trust*, *harm*, *human dignity*, *confidence*, *expectations*, *support*, *fairness* and *anxiety*. Each of these variables was measured on a self-reported scale of: 0 (strongly disagree) – 100 (strongly agree).

1. Blame: The drone is to blame for the action.
2. Trust: The drone can be trusted to take the correct actions in the future.
3. Harm: The actions of the drone caused harm.
4. Human dignity: The actions of the drone respect human dignity.
5. Confidence: I am confident that the drone will take the correct actions in the future.
6. Expectations: The actions of the drone are according to my expectations.
7. Support: I support the use of these type of drones by the military.
8. Fairness: The actions of the drone are fair.
9. Anxiety: The actions of the drone worry me.

The following constructs were used in pilot study 1 and 2, but dropped in the final study:

10. Blame commander: The commander is to blame for the action.
11. Trust commander: I trust that the commander will take correct actions in the future.
12. Harm commander: The actions of the commander caused harm.

3.4.4. Attention check

We added a question with an attention check in the middle of the dependent variable questions in which we asked to select the value 40 for that question. The attention check is often added to a questionnaire in Cognitive Psychology studies for example by Logg (2017) and people who fail to provide the correct

answer are excluded from the sample reasoning that they also fail to pay adequate attention to the other questions.

3.4.5. Demographic variables

We also added questions to collect demographic information on the respondents and added variables on: *age, gender, education level, occupation and nationality*. Next to these general demographic questions we asked if respondents have *experience with artificial intelligence*, if they *worked with drones* and if they *have been in a conflict zone*.

3.5. Analytical approach

This section describes the approach that we took to analyse the data. For each of the studies we performed the following steps in the data analysis.

3.5.1. Pre-process data

In the Qualtrics survey tool after closing the survey, a distinction is made between *responses in progress* and *complete responses*. Before starting the data analysis, the completion percentage of the responses in progress need to be checked and decided if they can be added to the complete responses. It turns out that when Amazon MTurk is used to distribute the survey, many respondents neglect to click on the last button which means that their responses are 99% complete, but these responses are valid and usable. If the respondent answered all the questions we added them to the list of complete responses. If respondents failed to answer all questions, for example the demographic questions, we deleted them from the sample. In the next pre-processing step, we deleted all respondents who failed the attention check that we build half way in the questionnaire. We asked if people could answer 40 to that question and we deleted all answers that fell outside the range of 38 and 42 which we assume is an accuracy mistake in placing the sliders.

3.5.2. Reliability analysis

After uploading the csv file in SPSS, we conducted a reliability test which is a measure to assess the internal consistency of a set of test items or scale. The reliability test checks if the given measurement is a consistent measurement of a concept. The construct to measure the internal consistency is Cronbach's Alpha (α) which should be over 0.7 to show internal consistency for the concept. For each of the studies the Cronbach's Alpha for the items of the agency construct is reported, because these items are used to measure the agency construct and should have internal consistency. It is also indicated if Cronbach's Alpha is improved if one of the items is removed. The reliability of the dependent variables is checked, but not reported, because these items are not designed to be used as one scale and it turned out that their internal consistency is lower than the threshold ($\alpha < 0.7$) as could be expected.

3.5.3. Principal Component Analysis (PCA)

In the next of the step in the data analysis a Principal Component Analysis (PCA) on the agency items was conducted. This step is a variable reduction technique aimed to reduce a larger set of variables into a smaller set, called 'principal components', that account for most of the variance in the original variables. We performed the PCA on the agency items to see if they could be aggregated to a single agency construct and report the variance and number of components as results. We also checked if the construct was influenced by the demographic variables, but will not report these results as the PCA indicates that all the demographic variables are distinct components that do not influence the agency construct.

3.5.4. Correlation analysis

The bivariate Pearson Correlation is a check to measure the strength and direction of the linear relationships between a pair of continuous variables. It produces a correlation coefficient r that lies between -1 and +1. A negative value indicates a negative relationship between the constructs, for example the larger the agency perception, the less the support. A positive value indicates a positive relationship, for example the larger the agency perception, the more blame. A correlation coefficient of zero indicates that no relationship between the constructs exists. We ran this analysis to check the correlation between the agency construct and the dependent variables and report the results.

3.5.5. Manipulation check on agency

The manipulation check examines if the agency construct varies between the scenarios and by this we checked if the hypotheses should be accepted or rejected. The check is done in two ways. First a graph is created, with the agency construct on the y-axis and scenarios on the x-axis, for a visual overview. The agency levels are expected to be different for the various conditions and this difference should be significant with $p < .05$. That is shown by the I shaped bars on top of the columns which show the standard error multiplied by 1 and indicates if the constructs overlap or not. If the bars overlap than the two groups are not significantly different.

The same result can be achieved by independent samples T-test which is a check for significant differences in the means of two distinct groups for which the subjects are randomly assigned so that the observed effect is the result of the treatment (in this case reading a specific scenario) and cannot be attributed to a different effect. We will show the graphs and report the results of the independent samples T-test for the agency construct between the Human Operated scenarios and scenarios with the different agency levels of Autonomous Weapons.

3.5.6. Dependent variables analysis

The analysis of the dependent variables is of an exploratory nature and the results are depicted by graphs. Each graph shows the mean of the dependent variable on the y-axis and the different scenarios on the x-axis to point out the differences between the scenarios. Although the analysis of the dependent variables is descriptive, we report the findings comparing between the different conditions that are significant at $p < .05$.

3.6. Pre-registration

The final study was pre-registered on the Open Science Framework⁸ which creates a time-stamped version of a project that cannot be deleted or edited. Although not required by scientific journals yet, it is often checked by reviewers if a study is pre-registered. We used the AsPredicted template in which 8 questions about the study are answered, for example if data already has been collected, which hypotheses the study tests and type of analyses you intend to perform. The pre-registration is placed under embargo until June 24, 2021. We preregistered the strong hypothesis we have about the agency perception, as we think we can replicate the findings based on the results of pilot study 2. For the dependent variables, we are not sure which mechanisms are at play and if we will be able to replicate our results therefore we registered an exploratory analysis on those results.

⁸ <https://osf.io/>

3.7. Sample

All surveys were created in Qualtrics and the pilot studies were distributed via crowdsourcing platform Amazon Mechanical Turk. We tested the survey with a small sample of 20 respondents and inspected the average completion time to see if we estimated the completion time correctly. We checked if the scenarios were evenly distributed, if people correctly checked the test question and if there were any remarks about the survey. Given that it all checked out fine, we decided to scale up the batch to 50 respondents per scenario which meant 500 participants for pilot study 1 and 700 for pilot study 2. The collection of the answers took about 4 hours and each respondent was paid USD 0.40.

The final study was distributed via the snowball method by email with an anonymous link to approximately 40 military colleagues who further distributed the survey. This method was used because we were not allowed by MIT and the Dutch MOD to collect any personal information, such as email or IP addresses. By using the snowball method, we had no guarantee on the number of respondents. We also published a news item on the internal Army website. The final study ran for 16 days from June 12 until June 27, 2017 and resulted in 327 responses of which 239 were complete valid responses and usable after the data pre-processing.

3.8. Methodological issues

In this section, several issues and concerns regarding the methodology will be addressed. These concerns could have implications for the internal and external validity of the study and if possible, some countermeasures are given which address the concerns.

3.8.1. Coding interviews

To get insight into the values of experts we conducted semi-structured interviews which is a qualitative research method. We used the Values Coding method to interpret the data and extract the values from the transcribed interviews. One of the concerns when coding qualitative data is that it is heuristic and no fixed formulas for coding exist. It also includes linking and providing meaning to the data (Saldaña, 2015). These characteristics imply that the Value Coding method is prone to bias in which the researcher uses their own framework and internal notions to interpret the data. To reduce the bias in applying the Values Coding method a second researcher will code the interviews independent from the first researcher. Some instruction on the method will be provided, but as information on the values and findings will be limited.

3.8.2. Randomized controlled experiments

One of the requirements of randomized controlled experiments is, as the name suggest, a random distribution of the respondents of the scenarios. The randomization limits confounding, enhances the internal validity of the research and is a crucial prerequisite for this type of study. One of the concerns is that the randomisation fails and that the distribution of respondents is skewed over the scenarios. One of the causes for skewed distribution is when the software fails to evenly distribute the respondents over the scenarios. As researchers, we can only check if the settings in the software are correct. Another cause for a skewed distribution could be that people see a certain scenario which they do not like or expect and quit the survey before completing. An example of this is when people expect to take a survey on Autonomous Weapons, but are assigned the Human Operated scenario and due to this decide to drop out of the survey. This prematurely exiting the survey is called selective attrition and it can result in a research bias, because the people who quit the survey are fundamentally different than the ones who take the survey. In the current research set-up, it is not possible to take countermeasures and it could be a threat to the internal validity of the study.

3.8.3. Amazon Mechanical Turk

The advantage of using Amazon MTurk as a survey distribution method is that a large group of respondents will participate in the survey and that will answer the questions seriously to maintain their workers status. This method also has some drawbacks because the workers on MTurk are primarily US based because in order to become a worker you will have to comply with US tax regulations even when you are a non-US resident. This could lead to a bias in the results which implies that the results are not generalizable beyond the US. Bonnefon et al. (2016) report a second concern with using MTurk in that participants might also not be representative for the US population, for example because many of the workers are students, and a third concern is that workers already are familiar with the materials. The first concern can be addressed by selecting a wide range of workers, so that the sample also includes people from outside the US. The second concern is inherent of the choice for MTurk and cannot be limited by countermeasures. The third concern is not very likely to occur because the scenarios are unique and have not been tested on MTurk before.

3.8.4. Snowball distribution final survey

The final study was distributed to military colleagues who forwarded the email to their network. Using this snowball distribution method has some drawbacks. The first is that participants could take the survey multiple times and as we are not allowed to store personal identifiable information we cannot check if this is the case. This could lead to a bias in the results if people see different scenarios and this could influence the answers on the questions. A second drawback is that colleagues take the survey first and tell the other participants of their findings before those take the survey. This also influences their answers. Thirdly, as researchers we cannot control who will take the survey so a potential risk would be that the survey gets send to someone outside the Ministry of Defence (MOD) by which the sample gets contaminated. Unfortunately, without storing personal identifiable information there are not many countermeasures that can be taken to address all three concerns.

4. Results

Hopefully people realize that computers will only do what they are programmed to do, and NOTHING else.

Respondent pilot study 1

This section describes the results of the value survey, that consists of an online questionnaire and expert interviews, and the randomized controlled experiments, comprised of two pilot studies and the final study. The values derived from the online questionnaire and interviews are combined in one overview. The results of the pilot and final studies are described using the structure of the data analysis steps (section 3.5) and the main results are listed in the conclusion of each study.

4.1. Value Survey

To get insight into which values people associate with Autonomous Weapons a survey and six expert interviews were conducted. A brief overview of the results is listed in this subsection.

4.1.1. Online survey

The value survey consisted of a questionnaire asking respondents first which values of BioEthics and Cheng and Fleischmann (2010) apply. In the final question, the participants were asked to provide at least one additional value that was not mentioned in the previous questions. In total 69 respondents took the survey and after removing the incomplete responses 57 complete responses were left. The results are depicted in graphs below.

This short questionnaire gives insight into which values people associate with Autonomous Weapons. The results from the BioEthics question shows that *non-maleficence* is most important, secondly it is *beneficence*, thirdly *justice* and lastly *autonomy*. The top five values based on the list of Cheng and Fleischmann (2010) are: *security, intelligence, responsibility, justice* and *social order*. The open questions reveal that people associate Autonomous Weapons mostly with *accountability* and *effectiveness*.

Bioethics values (Beauchamp & Walters, 1999)

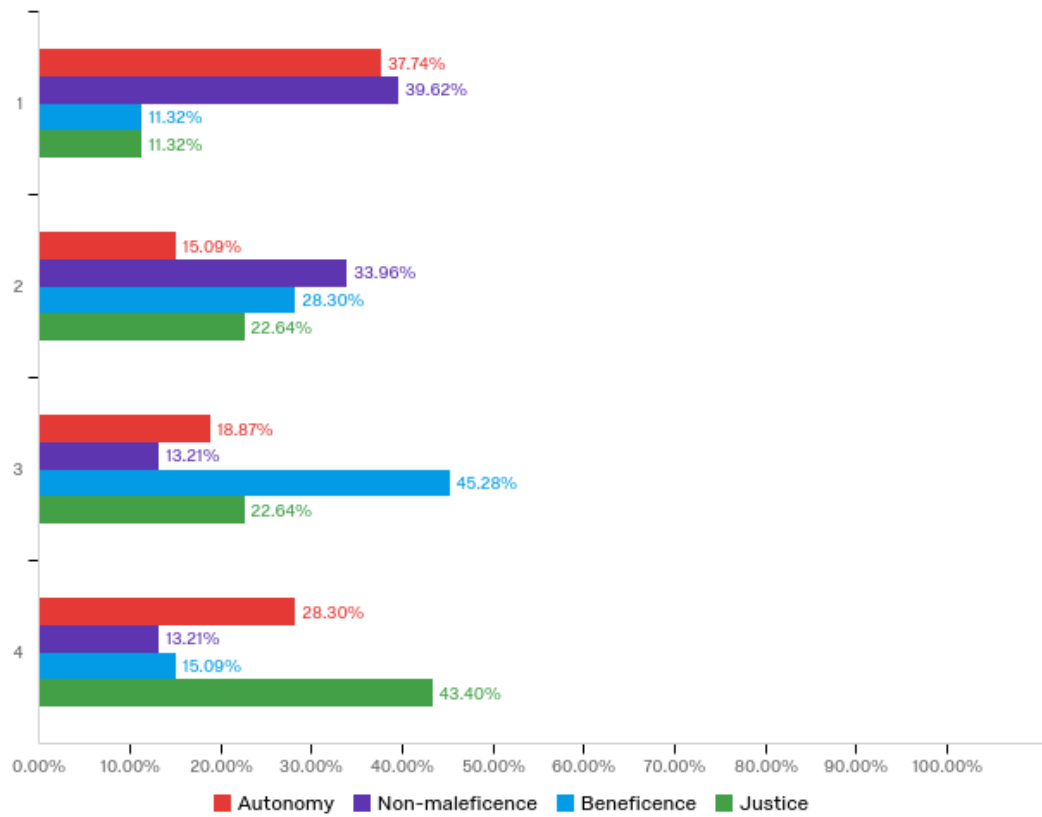


Figure 5 Results question 1 online value survey



Figure 6 Results question 3 online value survey

Universal human values (Cheng & Fleischmann, 2010)

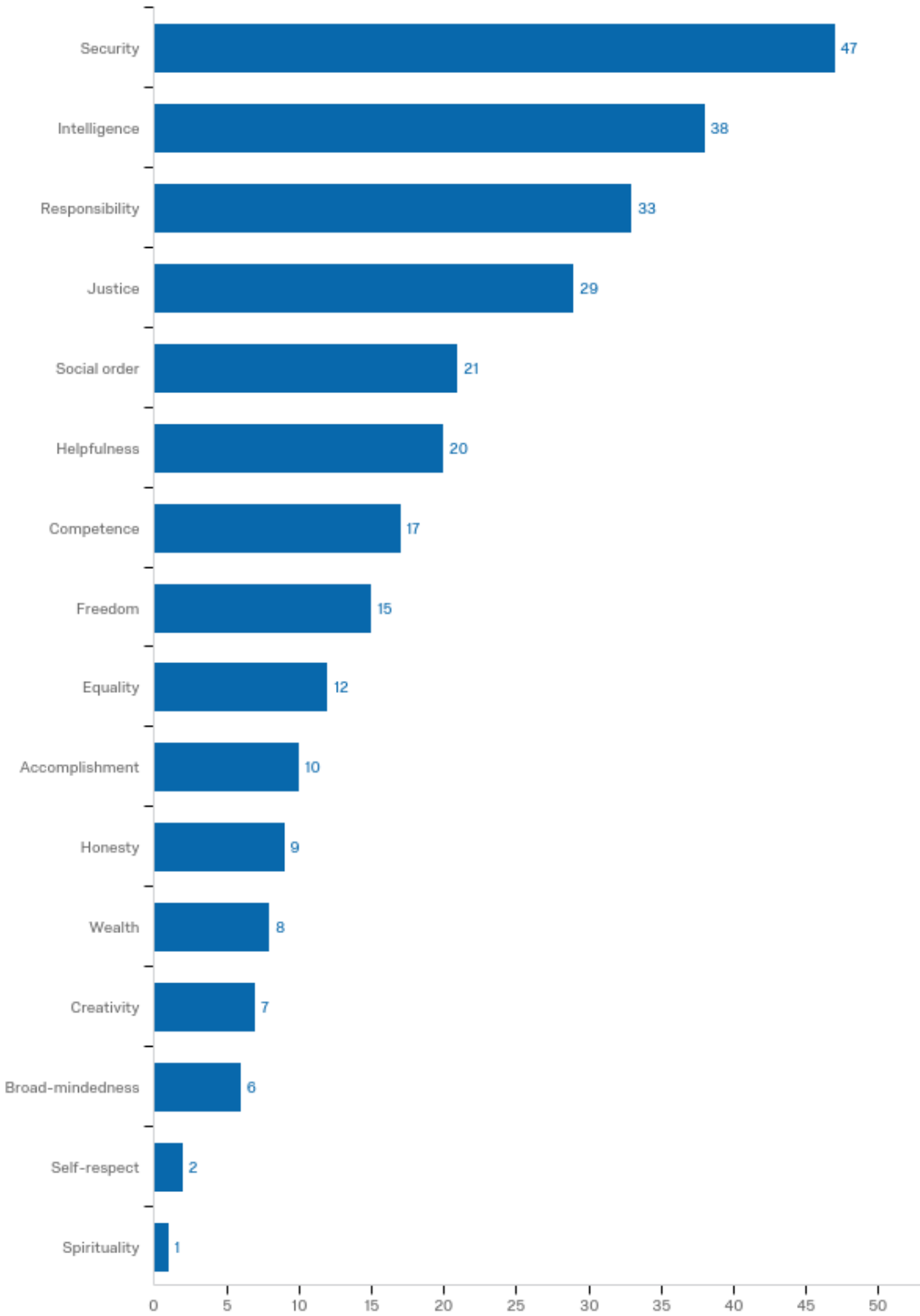


Figure 7 Results question 2 online value survey

4.1.2. Interviews

Six interviews were held with experts in the following fields:

Artificial Intelligence: prof. Toby Walsh (University of New South Wales & Data61).

Autonomous Weapons:

- dr. Peter Asaro (The New School & Campaign Ban Killer Robots);
- Mrs. Miriam Struyk (Program director Security & Disarmament PAX);
- dr. Michael Horowitz (University of Pennsylvania).

Military operations:

- LtCol Kremers (Dutch Army Headquarters);
- KLTZ van den Sande (MOD Headquarters).

Next to the discussion on values, the interviews provided rich background information on the history and positions of the different groups in the current debate on Autonomous Weapons, but also revealed that there is no consensus on a definition yet. The full transcriptions of the interviews can be found in Appendix F. Transcriptions interviews.

The interviews were coded by means of the ‘value coding’ method and based on the comparison of the value coding of the researcher and the second reviewer (Appendix H. Results coding process) the following list of 22 unique values are derived:

- | | |
|-----------------------------------|---------------------------|
| 1. Intervene at all times | 12. Distance |
| 2. Recall them | 13. Invincibility |
| 3. Minimize risk | 14. Determinism |
| 4. In control | 15. Inevitability |
| 5. Goodwill | 16. Unnecessary suffering |
| 6. Set boundaries | 17. Superfluous injury |
| 7. Right to be killed by a person | 18. Human dignity |
| 8. Ethical framework | 19. Accountability |
| 9. Meaningful human control | 20. Responsibility |
| 10. Predictability | 21. Defense |
| 11. Reflexivity | 22. Harm |

4.1.3. Conclusion value survey

We conclude on the value survey by comparing the values from the online survey and the interviews which results in a list of values (Table 6). Six values in this list are mentioned in both the value survey and the interviews. Based on the topics the six experts emphasised in the interviews, we selected the values that were mentioned most to be tested in pilot study 1. These values are *fairness*, *harm*, *human dignity* and *responsibility*.

Table 6 Overview values from online survey and interviews

Online survey	Interviews
Justice	
Autonomy	
Security	
Intelligence	
Social order	
Effectiveness	
Fairness	
Racism	
Control	In control
Beneficence	Goodwill
Accountability	Accountability
Responsibility	Responsibility
Safety	Defence
Non-maleficence	Harm
	Intervene at all times
	Recall them
	Minimize risk
	Set boundaries
	Right to be killed by a person
	Ethical framework
	Meaningful human control
	Predictability
	Reflexivity
	Distance
	Invincibility
	Determinism
	Inevitability
	Unnecessary suffering
	Superfluous injury
	Human dignity

4.2. Pilot study 1

In the first pilot study, we tried to test if we could manipulate the perception of agency of Human Operated drones and Autonomous Weapons. We were interested in how people perceive current technology compared to future technology. We created several conditions in which we varied the agency level either by providing extra information on how the drone deliberates about attacking the vehicle or by leaving this information intentionally out. We tested good outcome and bad outcome conditions, because we expected that this would have an impact on how the dependent variables, for example *support*, were viewed. In this study, we specifically asked the agency questions about the drone and not the Human Operator that is controlling it.

We tested 10 scenarios and aimed to get 50 respondents per condition and thereby a total of 500 respondents. The number of respondents per scenario ranged between 49 and 54. After the data pre-

processing and deleting the respondents who failed the attention check we were left with 510 complete responses.

Table 7 Scenarios of pilot study 1

		Agency status				
		Autonomous drone			Human operated drone	
		No agency	Low agency	High agency	Low agency	High agency
Outcome	No collateral damage (Good)	1	2	3	4	5
	Collateral damage (Bad)	6	7	8	9	10

4.2.1. Reliability analysis

For all five agency items $\alpha = 0.9$ which is well above the threshold of 0.7 and can only be improved if item SQ3_Moral rules is deleted as shown in Table 8. Deletion of other items will decrease the α value and reduce the internal consistency as shown in the final column (Cronbach's Alpha if item Deleted) of Table 8.

Table 8 Results reliability analysis agency items pilot study 1

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.900	.895	5

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
SQ1_Thought	167.77	13710.694	.841	.741	.857
SQ2_Goal setting	163.85	13440.002	.857	.775	.853
SQ3_Moral rules	183.39	17599.970	.448	.215	.933
SQ4_Act freely	163.22	13885.447	.824	.695	.861
SQ5_Achieve goals	157.21	13891.495	.794	.681	.868

4.2.2. Principal component analysis (PCA)

The PCA shows that *SQ1_Thought*, *SQ2_Goal setting*, *SQ4_Act freely* and *SQ5_Achieve goals* can be viewed as one construct which accounts for 71.75% variance in the original variables (Table 9). This is also seen in the scree plot that shows the eigenvalues of the components (Figure 8). Based on the component matrix we can observe that *SQ3_Moral rules* is a separate component compared to the other four agency items.

Table 9 Result PCA agency items pilot study 1

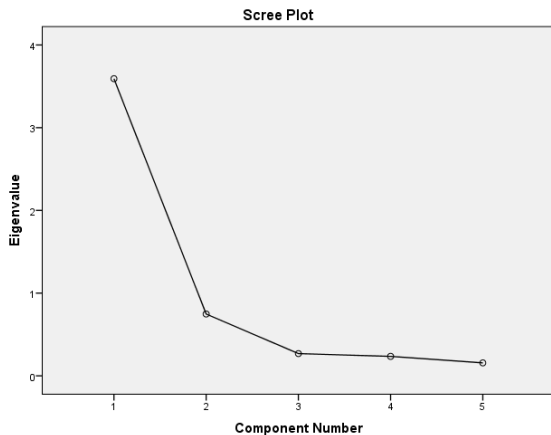


Figure 8 Scree plot PCA pilot study 1

Component Matrix^a

	Component		
	1	2	3
SQ2_Goal setting	.922		
SQ1_Thought	.911		
SQ4_Act freely	.899		
SQ5_Achieve goals	.882		
SQ3_Moral rules	.571	.819	

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

4.2.3. Correlation analysis

The correlation of the agency construct with the dependent variables is small and only for *DVQ1_Blame*, *DVQ2_CommanderBlame*, *DVQ4_CommanderTrust*, *DVQ6_CommanderHarm* and *DVQ8_Support* significant ($p < .01$) (Table 10). Some unexpected effects can be observed in the direction of relationships. For example, it is to be expected that the agency perception leads people to assign more *blame* to the drone ($r = .183$), but these results indicate that people also assign less *blame* to the *commander* ($r = -.239$). The same effect can be observed with the *harm* variable. The correlation analysis is not detailed enough to zoom in to these results, therefore this will be done in the analysis of the dependent variables.

4.2.4. Manipulation check on agency

The graph in Figure 9 shows that our agency manipulation works, because in the low agency condition for the Autonomous Weapon, people perceive less agency than in the high agency condition shown by the mean of the agency construct on the y-axis. In the neutral agency condition the agency perception is slightly higher than in the low agency condition, but still lower than the high agency condition. The independent samples T-test shows that the low agency AW and high agency AW conditions are distinct groups ($p = .000$), but that this is not the case for the neutral agency AW and the low agency AW ($p = .209$), because the difference between these groups is not significant which can also be seen by the overlap in the error bars for these conditions.

The Human Operated drone is perceived to have much lower agency than the Autonomous Weapon conditions. The agency manipulation works although the difference is less explicit than in the Autonomous Weapon scenario. The error bars of the agency level scenarios indicate that the Human Operated drone groups cannot be distinct which is confirmed by the independent sample T-test ($p = .125$).

The second graph (Figure 10) shows the mean of the agency perception over the agency levels for good and bad outcomes. The type of outcome caused by the weapon, being collateral damage or not, does not seem to lead to much differences in agency perception as the means differ 2 to 3 points (on a 100-point scale). The error bars also overlap in all agency level which is confirmed by the independent sample T-tests. In the neutral agency level condition, the significance between the bad outcome and good outcome group is $p = .816$, in the low agency condition $p = .641$ and in the high agency condition $p = .575$ which are all much higher than the $p < .05$ threshold.

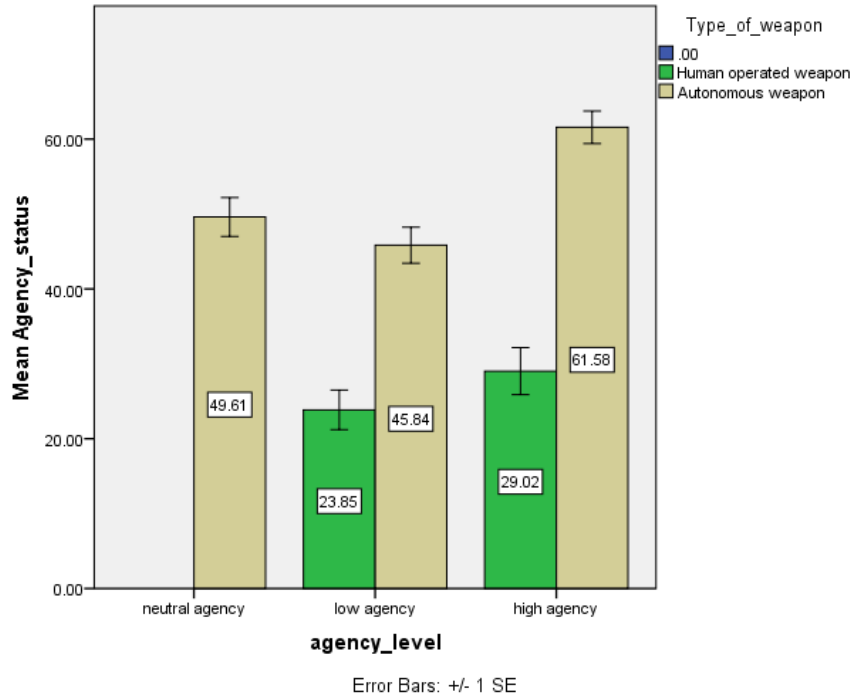


Figure 9 Mean value agency construct per Type of Weapon

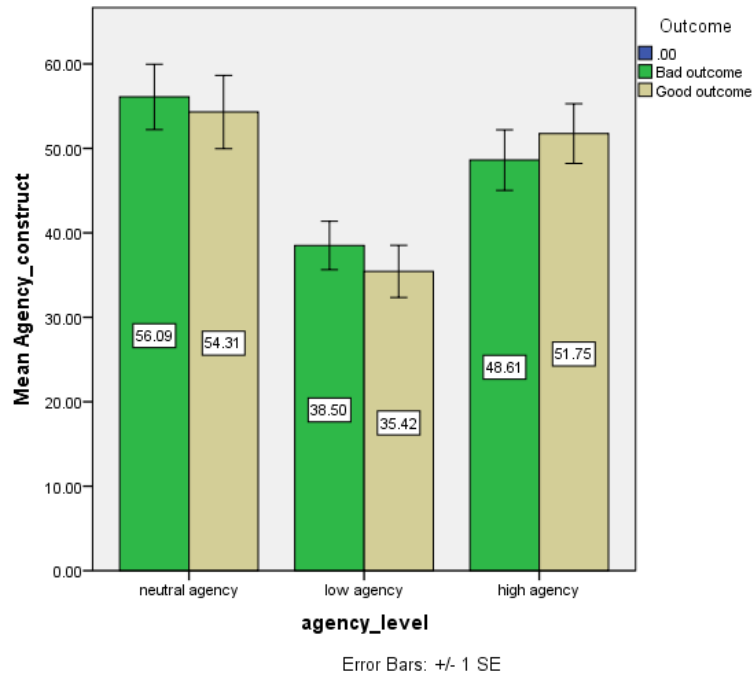


Figure 10 Mean value agency construct per condition per Outcome.

Table 10 Correlation matrix agency construct and dependent variables.

		Correlations									
		Agency_construct	DVQ1_Blame	DVQ2_Comm anderBlame	DVQ3_Trust	DVQ4_Comm anderTrust	DVQ5_Harm	DVQ6_Comm anderHarm	DVQ7_Uneasy	DVQ8_Support	DVQ9_Fairness
Agency_construct	Pearson Correlation	1	.183**	-.239**	.070	.176**	-.011	-.227**	.005	.101*	.087
	Sig. (2-tailed)		.000	.000	.114	.000	.798	.000	.914	.023	.050
	N	510	510	510	510	510	510	510	510	510	510
DVQ1_Blame	Pearson Correlation	.183**	1	-.101*	.020	.053	.198**	-.170**	.129**	-.120**	-.078
	Sig. (2-tailed)	.000		.023	.660	.232	.000	.000	.003	.006	.077
	N	510	510	510	510	510	510	510	510	510	510
DVQ2_CommanderBlame	Pearson Correlation	-.239**	-.101*	1	-.264**	-.330**	.264**	.692**	-.001	-.333**	-.036
	Sig. (2-tailed)	.000	.023		.000	.000	.000	.000	.982	.000	.422
	N	510	510	510	510	510	510	510	510	510	510
DVQ3_Trust	Pearson Correlation	.070	.020	-.264**	1	.635**	-.294**	-.262**	.169**	.674**	-.106*
	Sig. (2-tailed)	.114	.660	.000		.000	.000	.000	.000	.000	.016
	N	510	510	510	510	510	510	510	510	510	510
DVQ4_CommanderTrust	Pearson Correlation	.176**	.053	-.330**	.635**	1	-.171**	-.322**	.138**	.584**	.021
	Sig. (2-tailed)	.000	.232	.000	.000		.000	.000	.002	.000	.630
	N	510	510	510	510	510	510	510	510	510	510
DVQ5_Harm	Pearson Correlation	-.011	.198**	.264**	-.294**	-.171**	1	.361**	-.044	-.336**	.089*
	Sig. (2-tailed)	.798	.000	.000	.000	.000		.000	.323	.000	.046
	N	510	510	510	510	510	510	510	510	510	510
DVQ6_CommanderHarm	Pearson Correlation	-.227**	-.170**	.692**	-.262**	-.322**	.361**	1	-.038	-.322**	.041
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000		.394	.000	.351
	N	510	510	510	510	510	510	510	510	510	510
DVQ7_Uneasy	Pearson Correlation	.005	.129**	-.001	.169**	.138**	-.044	-.038	1	.127**	-.689**
	Sig. (2-tailed)	.914	.003	.982	.000	.002	.323	.394		.004	.000
	N	510	510	510	510	510	510	510	510	510	510
DVQ8_Support	Pearson Correlation	.101*	-.120**	-.333**	.674**	.584**	-.336**	-.322**	.127**	1	-.004
	Sig. (2-tailed)	.023	.006	.000	.000	.000	.000	.000	.004		.934
	N	510	510	510	510	510	510	510	510	510	510
DVQ9_Fairness	Pearson Correlation	.087	-.078	-.036	-.106*	.021	.089*	.041	-.689**	-.004	1
	Sig. (2-tailed)	.050	.077	.422	.016	.630	.046	.351	.000	.934	
	N	510	510	510	510	510	510	510	510	510	510

** . Correlation is significant at the 0.01 level (2-tailed).

* . Correlation is significant at the 0.05 level (2-tailed).

4.2.5. Dependent Variables analysis

The correlation analysis showed that only the dependent variables *blame*, *commander blame*, *commander trust*, *commander harm* and *support* were significant at a minimum of a $p < .05$ level. In the next step, we took a closer zoomed into these variables and due to the non-significance, we did not analyse the rest of the dependent variables. Looking more in detail we checked to see if there are any differences between the agency levels low, neutral and high and the Autonomous Weapon and Human Operated drone conditions.

Support variable

The most interesting observation is the large difference in *support* between good and bad outcomes (Figure 11) which is significant in all three conditions ($p < .05$). This substantial difference in the mean of *support* over the three conditions indicates that our outcome manipulation works. Another interesting observation is that the *support* goes down as agency levels go up, but the differences in the mean are small (< 10 points on a 100-point scale) and the I shaped bars on top with the standard error overlap indicating that these groups are not significantly different, when comparing them in either the good outcome or the bad outcome scenarios. The second graph shows that the support is higher in the Human Operated conditions compared to the Autonomous Weapon conditions, but also here the differences in the group are small and not significant as the error bars overlap.

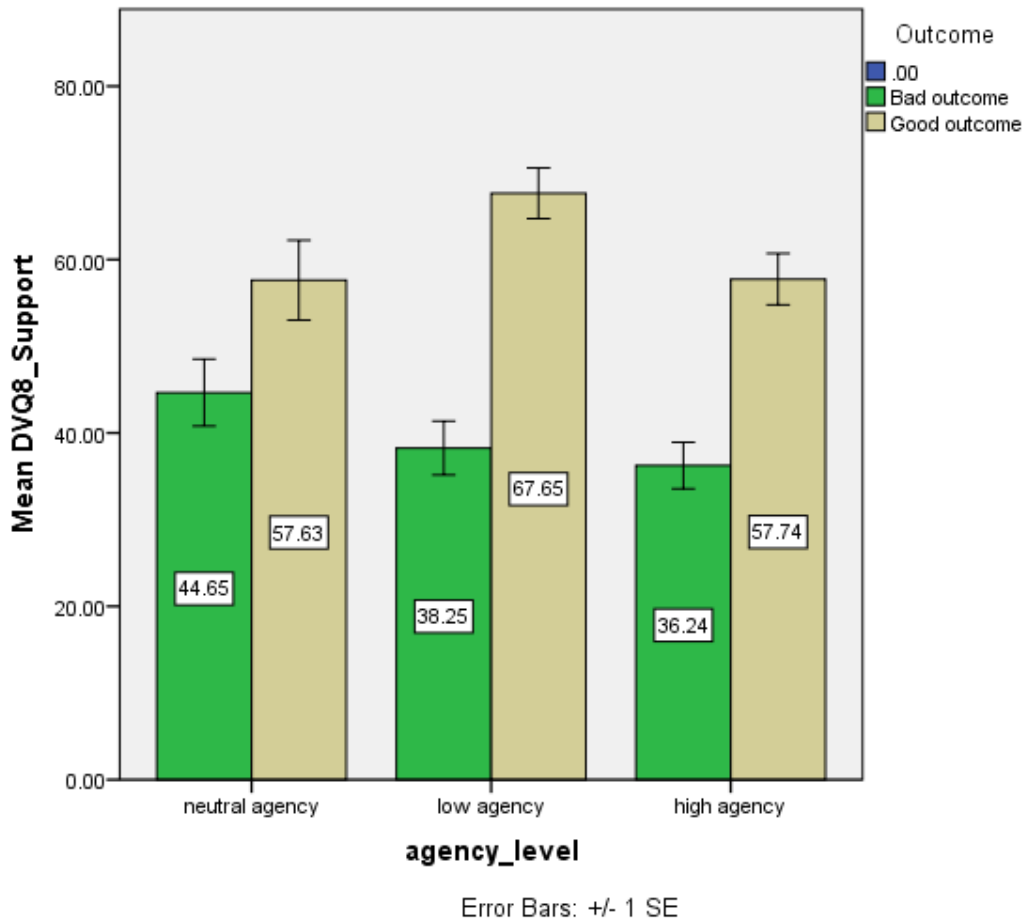


Figure 11 Mean value support variable per condition per Outcome

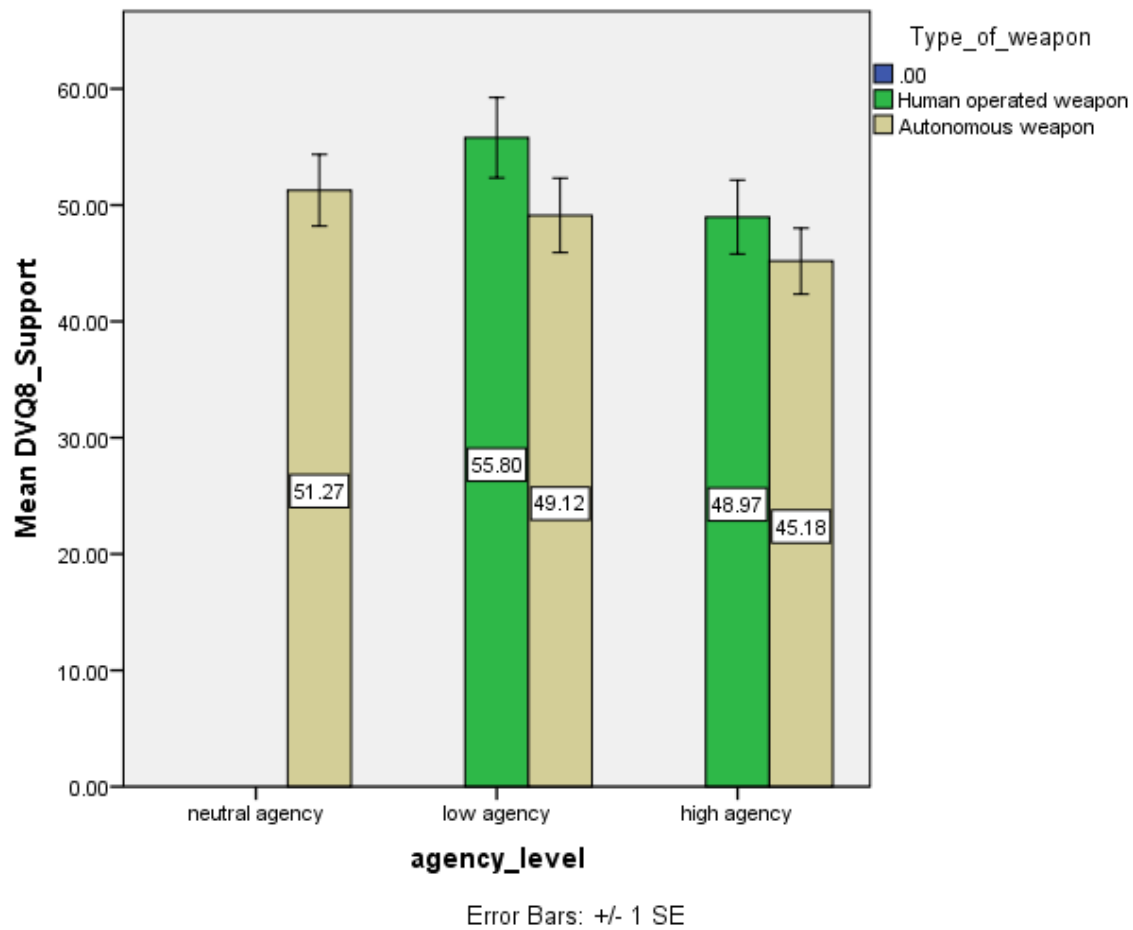


Figure 12 Mean value support variable per condition per Type of Weapon

Trust variable

From the graph in Figure 12 it can be observed that people are less likely to *trust* an Autonomous Weapon to take correct actions in the future after a bad outcome than a human operator who takes an action with the same bad outcome and all results are significant. This effect is magnified in low agency condition where we see almost no difference between *trust* in human after good or bad outcome and large difference in *trust* in AW (Figure 13). However, the univariate analysis showed that the interaction is not significant: *Type_of_weapon * outcome* $p = .120$ so this means that we cannot draw any conclusions at this point of the study, but it is an interesting finding to further investigate.

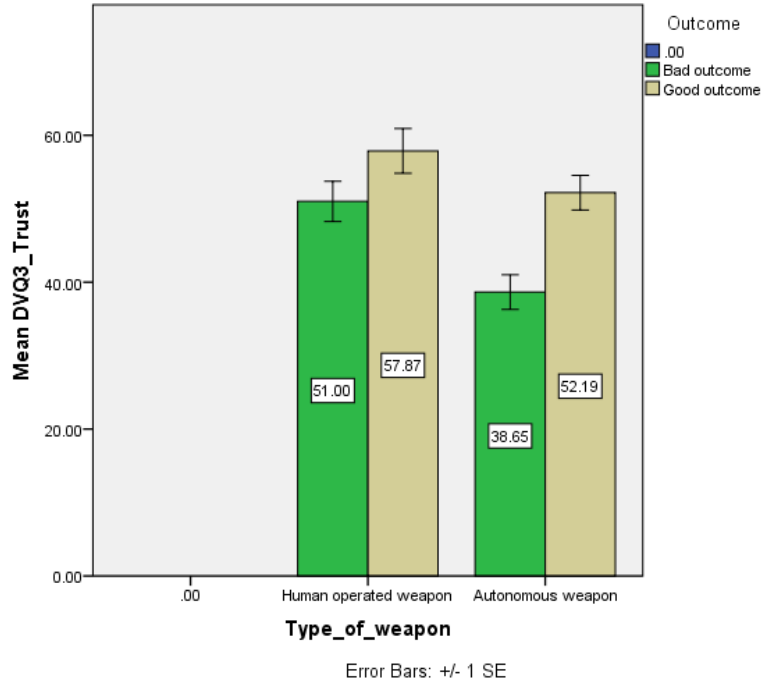


Figure 13 Mean value trust variable per Type of Weapon per Outcome

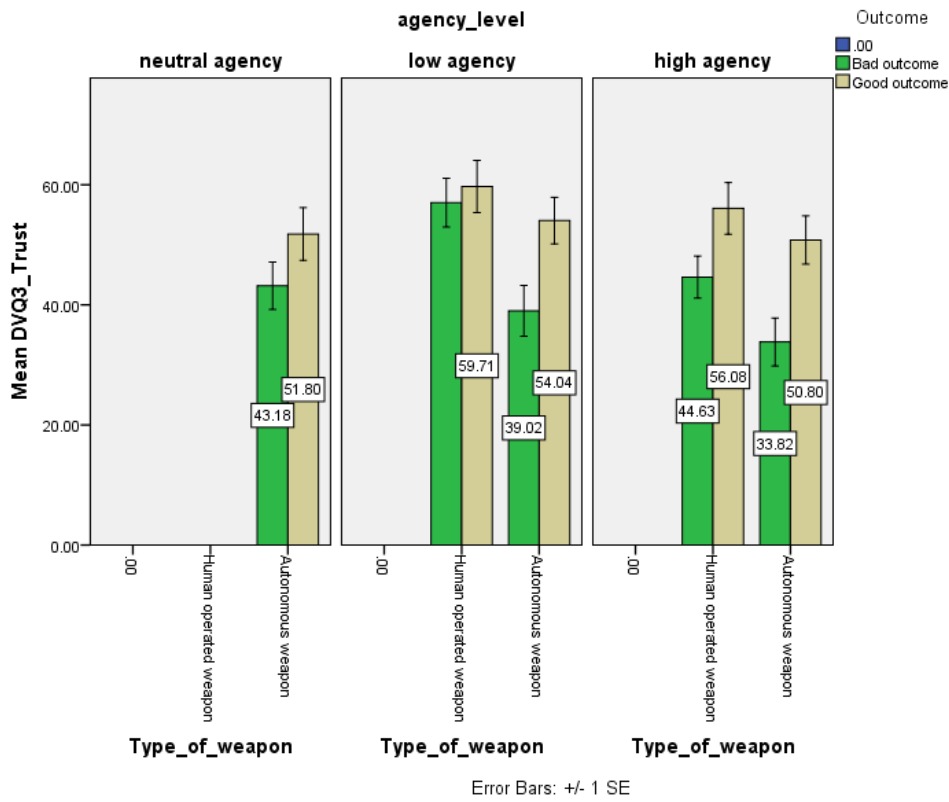


Figure 14 Mean value trust variable per condition per Outcome

Blame and harm variables

To study the chain of responsibility the mean average of the variables *blame*, *commander blame*, *harm* and *commander harm* are depicted in the same graph (Figure 15). All results of the blame variables (Figure 15), except for the neutral agency Autonomous Weapon condition, are significant. These results show that in low agency conditions the *blame* is shifted to the *commander* and in high agency conditions the *blame* is put on the drone.

There is one striking observation in that the low agency condition has a surprisingly high amount of causal responsibility for the *harm*. This pattern holds for Human Operated drones and Autonomous Weapons equally (Figure 16). Contrary to the *blame* variable, the drone is seen as doing much *harm* in both high as low agency conditions, but in a low agency condition the *harm* is shifted almost equally to the *commander*. This means that they are both causing *harm*, but in high agency conditions it is much less transferred to the *commander* and the drone is seen as causing the *harm*. However, the difference between the variables *harm* and *commander harm* in the low agency condition of both the Human Operated as the Autonomous Weapon scenarios is not significant therefore we must caution with drawing conclusions on the observed effect.

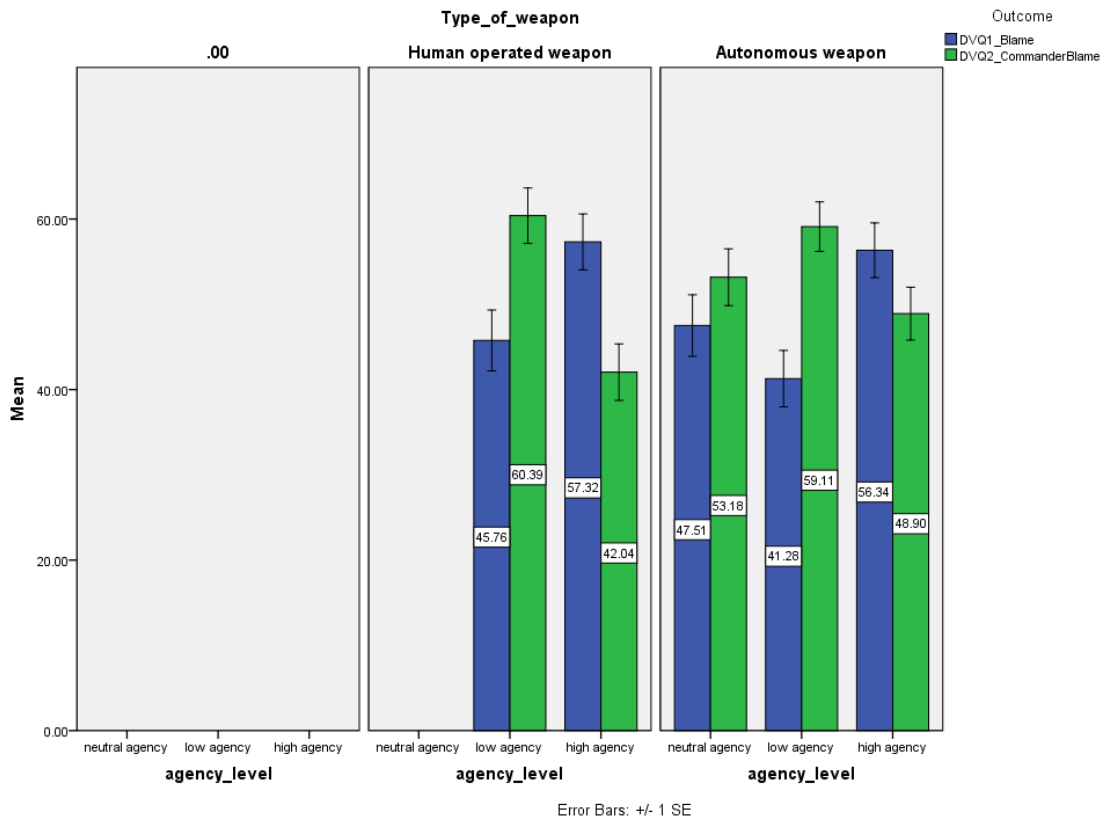


Figure 15 Mean value blame and commander blame variables per condition and Type of Weapon

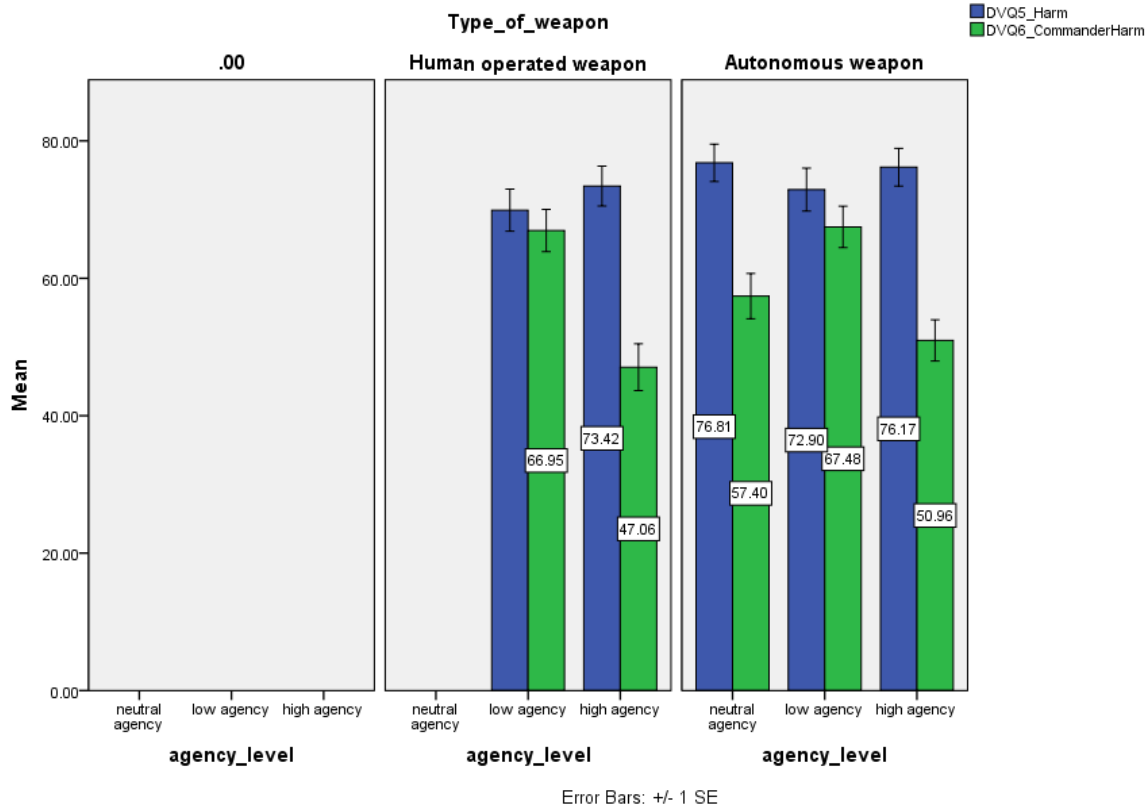


Figure 16 Mean value harm and commander harm variables per condition and Type of Weapon

4.2.6. Conclusion pilot study 1

The results of pilot study 1 lead to the following conclusions:

- Based on the results of the PCA and reliability analysis of the agency items, we decided to drop *SQ3_Moral rules* from the agency construct which improves the reliability of the construct to $\alpha = 0.933$. In the next study, we will use the agency construct that contains four items being *thought, goal setting, act freely* and *achieve goals*.
- The results of the T-tests indicate that there is a difference in the mean of the agency perception between the Autonomous Weapon and Human Operated conditions. There is also much difference between the high agency and low/neutral agency scenarios. However, between low and neutral agency there was not much difference observed. Therefore, we decided to drop the neutral agency condition in the next study and only test the low agency and high agency conditions.
- The most explicit observation in the *support* variable is that there is a big difference in support between good and bad outcomes (Figure 11). However this is what to be expected based on literature, because James Igoe Walsh (2015) and Kreps (2014) found that the general public has more *support* for drone strikes if there is no collateral damage compared to those that have collateral damage, but it shows that our outcome manipulation works and we will use this in the next study.
- In the analysis of the *trust* variable (Figure 13) we found that people are less likely to *trust* an Autonomous Weapon to take correct actions in the future after a bad outcome than a human operator who takes an action with the same bad outcome. This might indicate algorithm aversion for which algorithms are punished more than humans who make mistakes (Dietvorst, 2016) and we used these findings for the next pilot study to further study the algorithm aversion effect.

- We found that that in the low agency condition a surprisingly high amount of causal responsibility for the *harm* can be seen and that this pattern holds for Human Operated drones and Autonomous Weapons equally (Figure 16). We also found that the drone is seen as doing much *harm* in both high as low agency conditions, but in a low agency condition the harm is shifted almost equally to the *commander*. Although these findings are not significant, they are interesting and require more study therefore we kept these questions in pilot study 2.

4.3. Pilot study 2

In pilot study 2 we tried to get a better understanding of the algorithm aversion effects so we tested scenarios describing that the Autonomous Weapon could learn from its mistakes (3 + 10), was trained on a large data set and that it understands its mission (4 + 11), that it could be sometimes unpredictable (5 + 12) and a scenario that includes all of these aspects (6 + 13). We tested both on good outcome and bad outcome conditions and added a scenario for the Autonomous Weapon with low agency (1 + 8) to check if our agency manipulation works and a Human Operated scenario (7 + 14) in which we asked about the human operator instead about the drone which is different from pilot study 1.

We tested at total of 14 scenarios and aimed for 50 respondents per scenario which means 700 responses in total. The number of respondents per scenario ranged between 47 and 54. After the data pre-processing and deleting the respondents who failed the attention check we got 709 complete responses.

Table 11 Scenarios of pilot study 2

		Aspects						
		Autonomous drone						HO
		Level 1 - Low agency	Level 2 (high agency + no extra)	Level 3 (high agency + learning ability)	Level 4 (high agency + understanding procedures)	Level 5 (high agency + unpredictability)	Level 6 (high agency + all aspects)	Human operated drone
Outcome	Collateral damage (Bad)	1	2	3	4	5	6	7
	No collateral damage (Good)	8	9	10	11	12	13	14

4.3.1. Reliability analysis

For all four agency items $\alpha = 0.914$ which is well above the threshold of 0.7 and cannot be improved by deleting an item. Deletion of other items will decrease the α value and reduce the internal consistency.

Table 12 Results reliability analysis pilot study 2

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.914	.914	4

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Thought	208.4827	6427.511	.801	.647	.890
Goal_setting	204.9885	6499.803	.834	.695	.877
Act_freely	203.4020	6782.379	.804	.646	.888
Achieve_goals	199.4975	7114.019	.781	.614	.897

4.3.2. Principal component analysis (PCA)

The PCA shows that *Thought*, *Goal setting*, *Act freely* and *Achieve goals* can be viewed as one construct which accounts for 79.60 % variance in the original variables (Table 13). This is also seen in the scree plot that shows the eigenvalues of the components (Figure 17). The PCA shows that the four items; *thought*, *goal setting*, *act freely* and *achieve goals*, can be viewed as the underlying components of the agency construct (Table 13).

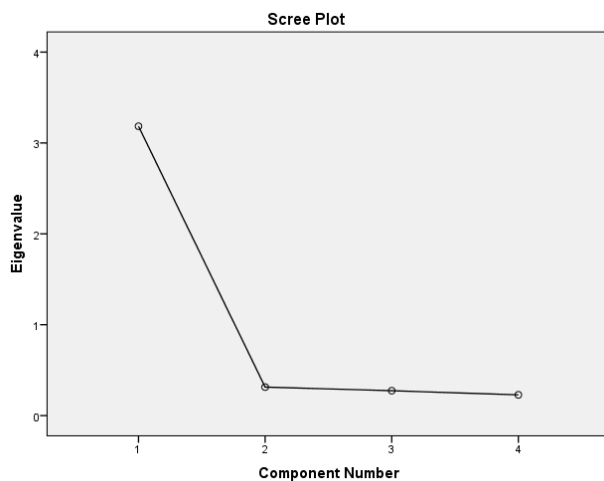


Figure 17 Scree plot PCA pilot study 2

Table 13 Results PCA pilot study 2

	Component	
	1	2
Goal_setting	.910	
Act_freely	.892	
Thought	.889	
Achieve_goals	.877	.446

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

4.3.3. Correlation analysis

The correlation of the agency construct with the dependent variables is small, but significant for all constructs ($p < .05$) (Table 14). The same effects in direction for the *harm* and *blame* variable as in pilot study 1 can be observed in this study. There is a negative relationship between the agency construct and the level of uneasy meaning that the when the agency perception increases people indicate that they feel more *anxiety*, but this effect is very small ($r = -.076$). The correlation analysis is not detailed enough to zoom in to these results, therefore this will be done in the analysis of the dependent variables.

4.3.4. Manipulation check on agency

The graph in Figure 18 shows that there is much difference between the high agency scenarios and the low agency scenario and this difference is significant. The difference in the low agency and high agency conditions confirm that our agency manipulation works. The agency of the human operator and high agency drones are at the same level (in this study we specifically asked about the human operator in this pilot study compared to the drone in the previous study). There is some difference between the good outcome and bad outcome conditions and although these effects are largest for the conditions in which we zoomed into the algorithm aversion, the difference is not very large (a difference in mean of 8 points on a 100-point scale). Also, the error bars between the good outcome and bad outcome conditions for all scenarios overlap indicating that these are not distinct groups. The only distinct and significant difference in groups ($p < .05$) is between the low agency Autonomous Weapon condition and the other scenarios. This is also confirmed by the level of significance of the independent samples T-test which we chose not to report for the graph below for the sake of clarity.

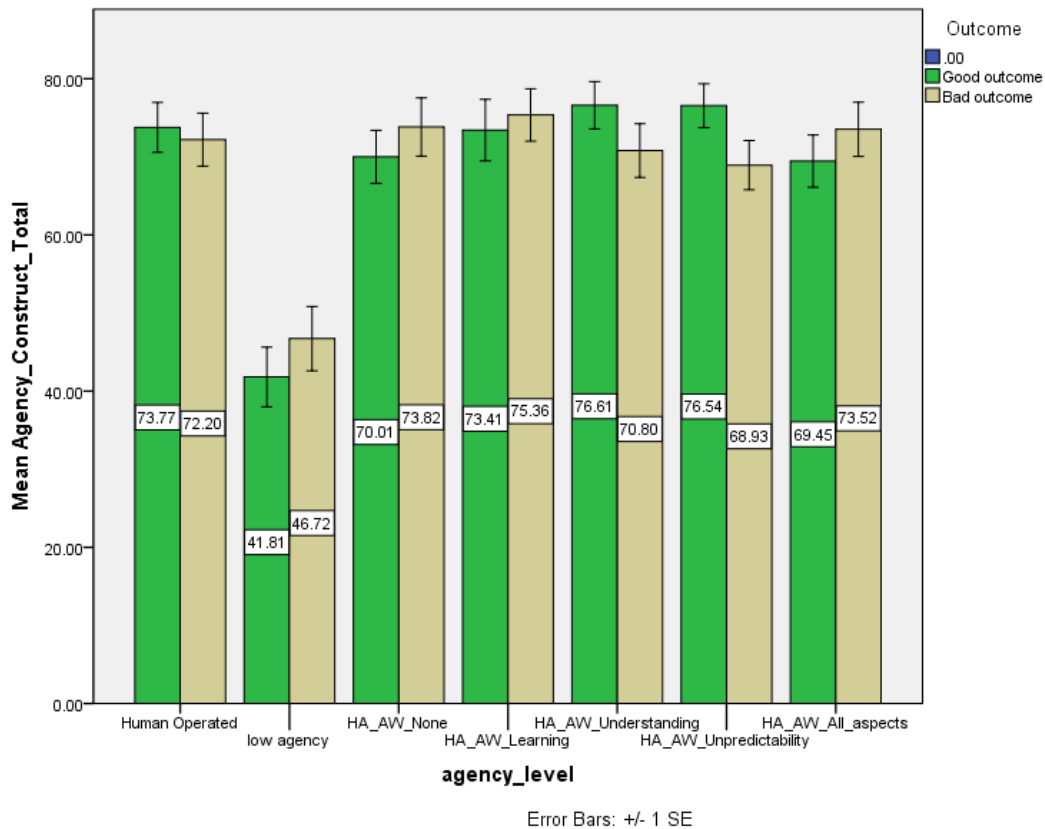


Figure 18 Mean value agency construct per condition per Outcome

Table 14 Correlation matrix agency construct and dependent variables pilot study 2

		Correlations												
		Agency_Construct_Total	DVQ1_Blame	DVQ2_CommanderBlame	DVQ3_Trust	DVQ4_CommanderTrust	DVQ5_Harm	DVQ6_CommanderHarm	DVQ7_Dignity	DVQ8_Confidence	DVQ9_Expectations	DVQ10_Support	DVQ11_Fair	DVQ12_Uneasy
Agency_Construct_Total	Pearson Correlation	1	.271**	-.267**	.205**	.217**	.124**	-.198**	.102**	.209**	.146**	.164**	.142**	-.076*
	Sig. (2-tailed)		.000	.000	.000	.000	.001	.000	.007	.000	.000	.000	.000	.043
	N	708	708	708	708	708	708	708	708	708	708	708	708	708
DVQ1_Blame	Pearson Correlation	.271**	1	-.069	-.028	-.046	.258**	-.023	-.056	-.048	-.137**	-.118**	-.092*	.180**
	Sig. (2-tailed)	.000		.065	.464	.226	.000	.534	.138	.198	.000	.002	.014	.000
	N	708	708	708	708	708	708	708	708	708	708	708	708	708
DVQ2_CommanderBlame	Pearson Correlation	-.267**	-.069	1	-.287**	-.278**	.152**	.691**	-.206**	-.300**	-.205**	-.250**	-.275**	.312**
	Sig. (2-tailed)	.000	.065		.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	N	708	708	708	708	708	708	708	708	708	708	708	708	708
DVQ3_Trust	Pearson Correlation	.205**	-.028	-.287**	1	.653**	-.281**	-.224**	.595**	.901**	.533**	.726**	.722**	-.616**
	Sig. (2-tailed)	.000	.464	.000		.000	.000	.000	.000	.000	.000	.000	.000	.000
	N	708	708	708	708	708	708	708	708	708	708	708	708	708
DVQ4_CommanderTrust	Pearson Correlation	.217**	-.046	-.278**	.653**	1	-.185**	-.228**	.432**	.635**	.425**	.631**	.552**	-.458**
	Sig. (2-tailed)	.000	.226	.000	.000		.000	.000	.000	.000	.000	.000	.000	.000
	N	708	708	708	708	708	708	708	708	708	708	708	708	708
DVQ5_Harm	Pearson Correlation	.124**	.258**	.152**	-.281**	-.185**	1	.311**	-.374**	-.301**	-.217**	-.253**	-.324**	.395**
	Sig. (2-tailed)	.001	.000	.000	.000	.000		.000	.000	.000	.000	.000	.000	.000
	N	708	708	708	708	708	708	708	708	708	708	708	708	708
DVQ6_CommanderHarm	Pearson Correlation	-.198**	-.023	.691**	-.224**	-.228**	.311**	1	-.183**	-.226**	-.203**	-.235**	-.248**	.289**
	Sig. (2-tailed)	.000	.534	.000	.000	.000	.000		.000	.000	.000	.000	.000	.000
	N	708	708	708	708	708	708	708	708	708	708	708	708	708
DVQ7_Dignity	Pearson Correlation	.102**	-.056	-.206**	.595**	.432**	-.374**	-.183**	1	.588**	.415**	.533**	.598**	-.509**
	Sig. (2-tailed)	.007	.138	.000	.000	.000	.000	.000		.000	.000	.000	.000	.000
	N	708	708	708	708	708	708	708	708	708	708	708	708	708
DVQ8_Confidence	Pearson Correlation	.209**	-.048	-.300**	.901**	.635**	-.301**	-.226**	.588**	1	.553**	.723**	.725**	-.649**
	Sig. (2-tailed)	.000	.198	.000	.000	.000	.000	.000	.000		.000	.000	.000	.000
	N	708	708	708	708	708	708	708	708	708	708	708	708	708
DVQ9_Expectations	Pearson Correlation	.146**	-.137**	-.205**	.533**	.425**	-.217**	-.203**	.415**	.553**	1	.511**	.621**	-.509**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000	.000		.000	.000	.000
	N	708	708	708	708	708	708	708	708	708	708	708	708	708
DVQ10_Support	Pearson Correlation	.164**	-.118**	-.250**	.726**	.631**	-.253**	-.235**	.533**	.723**	.511**	1	.685**	-.651**
	Sig. (2-tailed)	.000	.002	.000	.000	.000	.000	.000	.000	.000	.000		.000	.000
	N	708	708	708	708	708	708	708	708	708	708	708	708	708
DVQ11_Fair	Pearson Correlation	.142**	-.092*	-.275**	.722**	.552**	-.324**	-.248**	.598**	.725**	.621**	.685**	1	-.656**
	Sig. (2-tailed)	.000	.014	.000	.000	.000	.000	.000	.000	.000	.000	.000		.000
	N	708	708	708	708	708	708	708	708	708	708	708	708	708
DVQ12_Uneasy	Pearson Correlation	-.076*	.180**	.312**	-.616**	-.458**	.395**	.289**	-.509**	-.649**	-.509**	-.651**	-.656**	1
	Sig. (2-tailed)	.043	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	
	N	708	708	708	708	708	708	708	708	708	708	708	708	708

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

4.3.5. Dependent variables analysis

The correlation between the agency construct and the dependent variables are all significant ($p < .05$) therefore the most striking findings for each of the dependent variables is described below.

Blame

The most notable finding for the *blame* variable is that people assign less *blame* in the low agency condition compared to the high agency Autonomous Weapon and the Human Operator conditions. For these conditions, the difference is significant ($p < .05$). Some effects of the algorithm aversion can be seen in the learning and understanding scenario where people assign less *blame* when the Autonomous Weapons does not make a mistake and more *blame* when it makes a mistake with a bad outcome. However, these conditions show an overlap in error bars and are not significant.

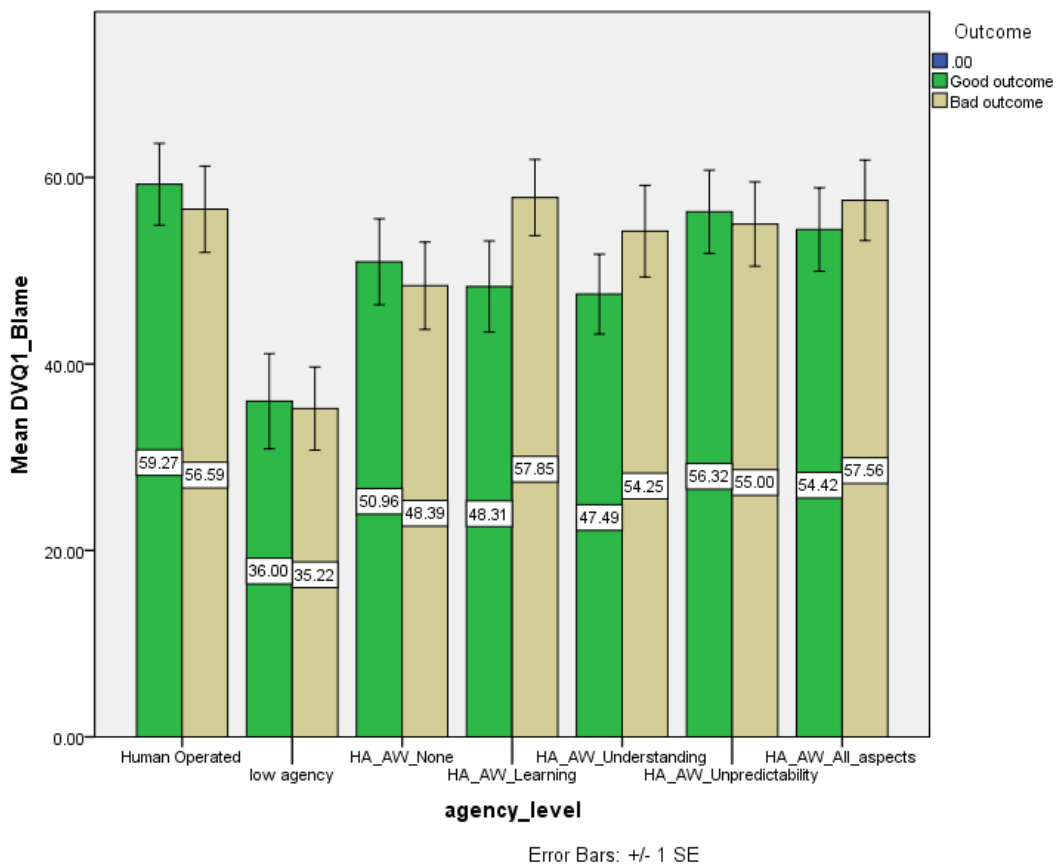


Figure 19 Mean value blame variable per condition per outcome

Commander Blame

The most striking observation in the *commander blame* variable is that the *commander* is *blamed* a lot less when the Autonomous Weapon makes a mistake in the learning and unpredictability conditions and this is significant ($p < .05$). It might be the case that the *blame* is shifted to the drone as people expect it to be able to learn from its mistakes or it behave unpredictable and that the *commander* is not to *blame* for that. Also, in the low agency condition the *commander* is *blamed* a bit more than in the other conditions.

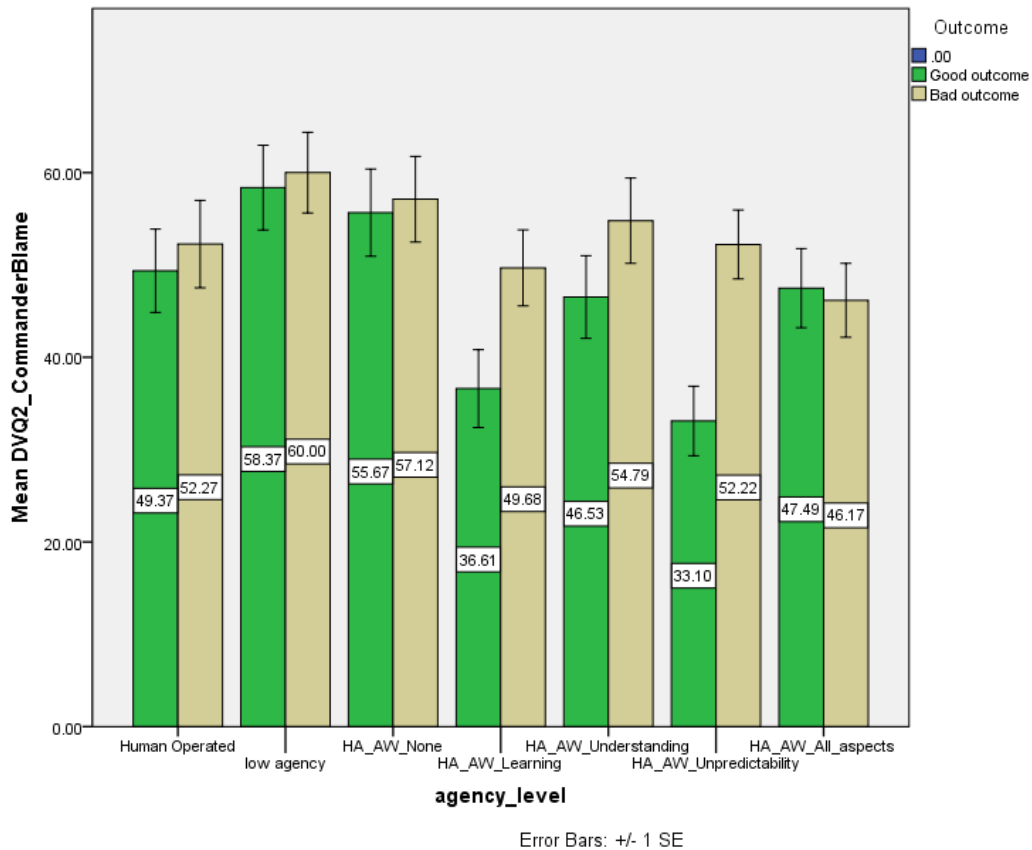


Figure 20 Mean value commander blame variable per condition per outcome

Trust

Over all the conditions it can be observed that people have more trust that the Autonomous Weapon or Human Operator will take correct actions in the future after a good outcome than a bad outcome. Except for the low agency condition, the difference in *trust* in the good and bad outcome conditions is significant ($p < .05$). People *trust* the Human Operator more than the Autonomous Weapon, especially in the low agency condition and the difference between these groups is significant ($p < .05$).

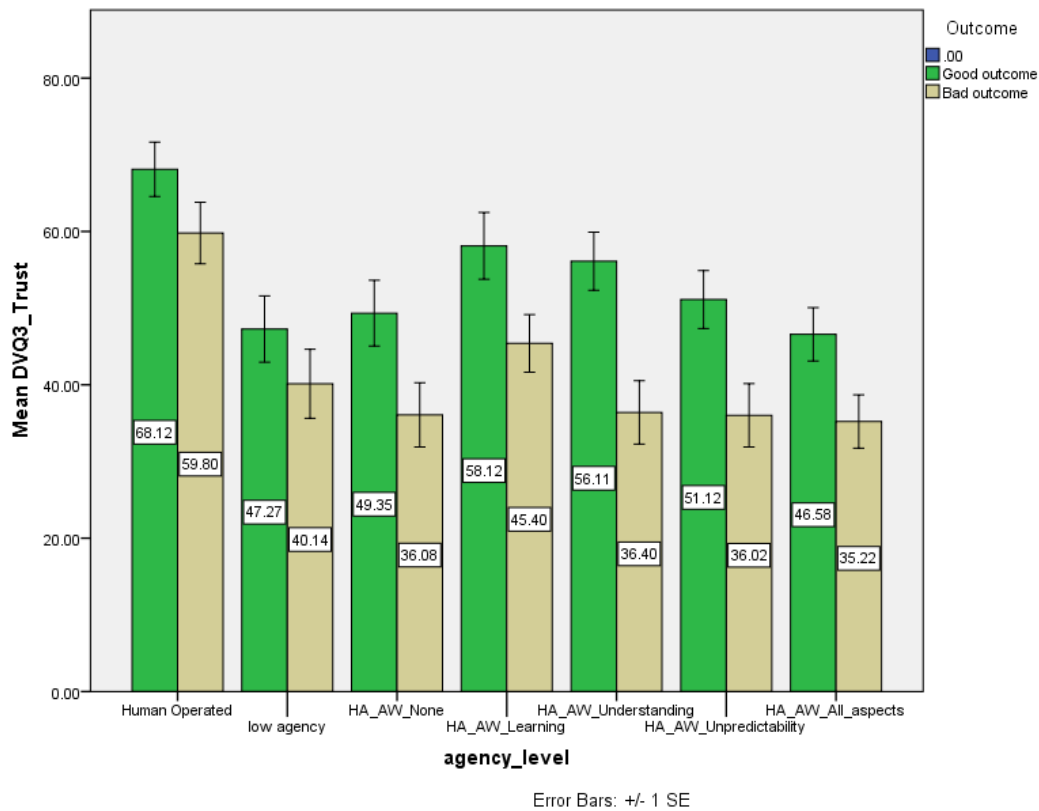


Figure 21 Mean value trust variable per condition per outcome

Commander Trust

Overall people *trust* that the *commander* will take correct actions more in the good outcome conditions compared to the bad outcome conditions. The *trust* in the *commander* is lowest in the low agency condition and high agency condition with no extra information and the difference between these groups is significant ($p < .05$). Knowing that the Autonomous Weapon has understanding capabilities results in higher trust in the bad outcome condition that is significant ($p < .05$).

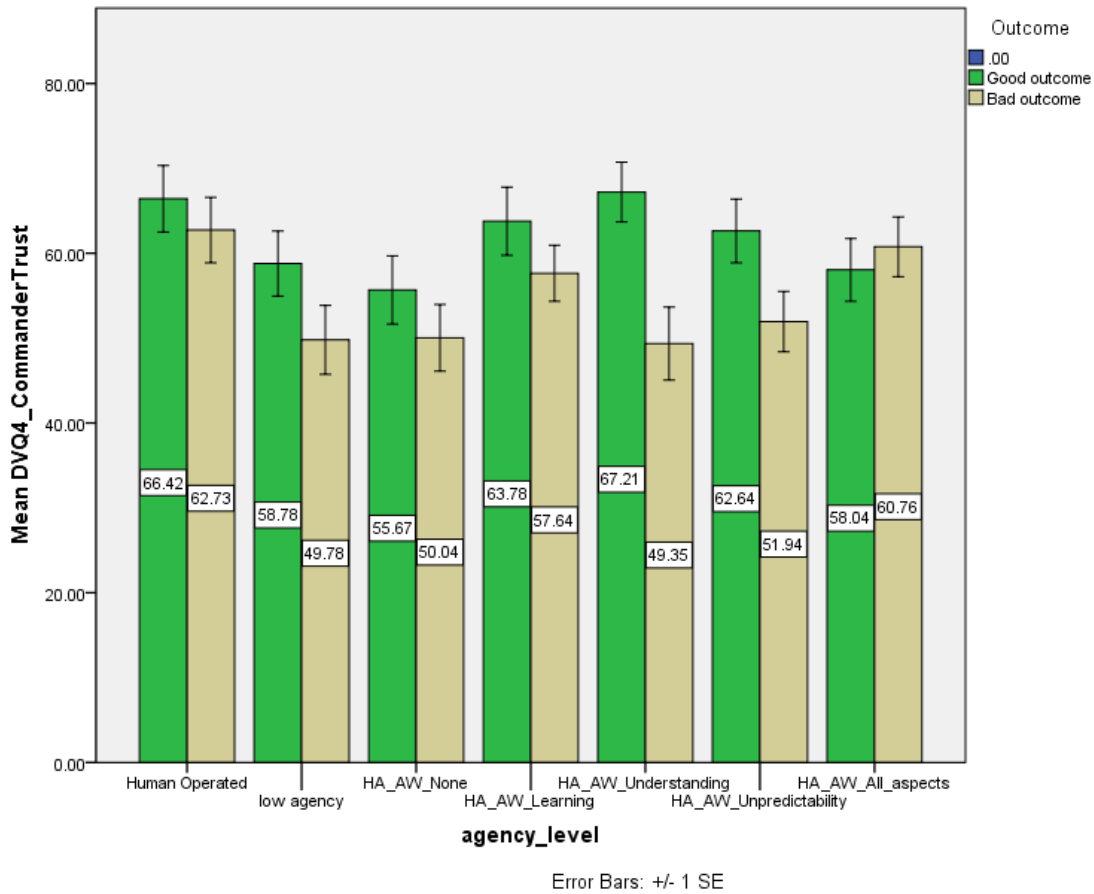


Figure 22 Mean value commander trust variable per condition per outcome

Harm

In the bad outcome conditions people perceive more *harm* than in the good outcome conditions which could be expected and the difference in nearly all conditions is significant ($p < .05$). Most notable in this graph is that when the Autonomous Weapon behaves unpredictable, the type of outcome does not seem to matter and the perception of harm is equal, but the difference between both outcome groups is not significant.

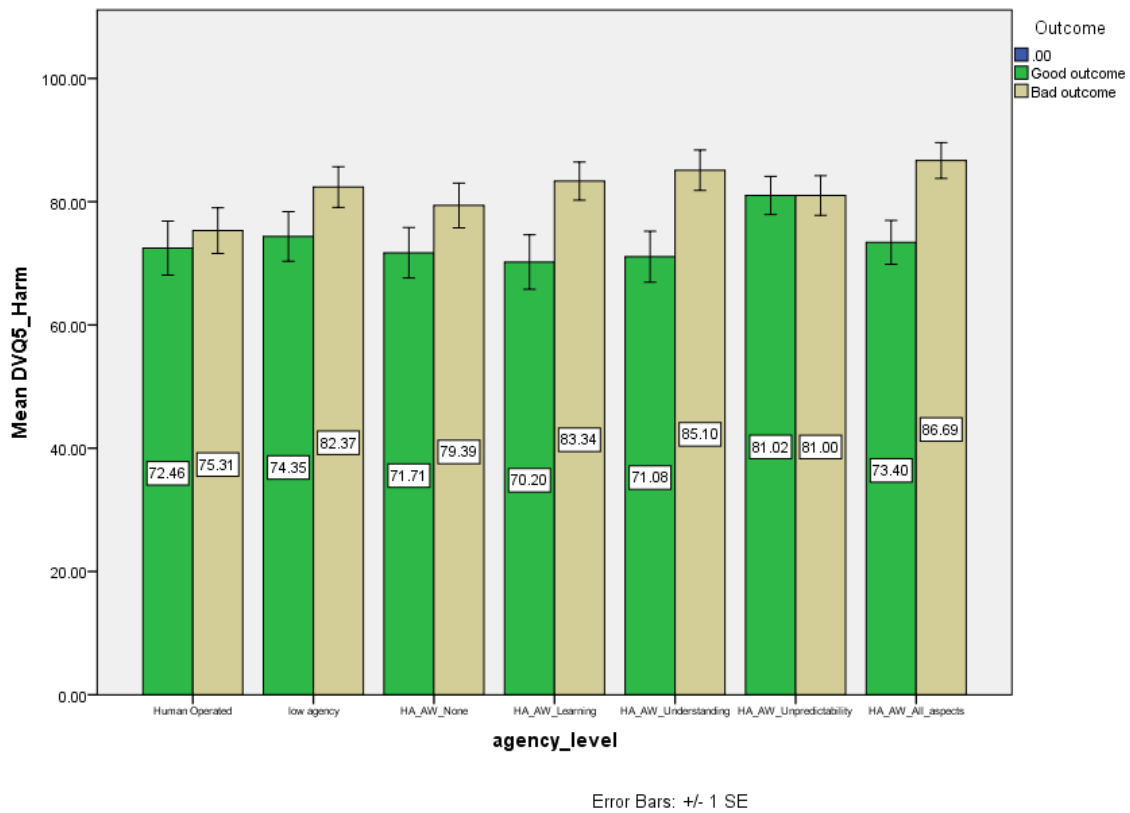


Figure 23 Mean value harm variable per condition per outcome

Commander Harm

In the bad outcome conditions people perceive the *commander* doing more *harm* than in the good outcome conditions which could be expected. Most striking is the large difference in the learning condition between the good and bad outcome where the commander is attributed much less harm in the good outcome condition which is significant ($p < .05$). The differences between good and bad outcome conditions between in the Human Operator and high agency unpredictability scenarios are also significant ($p < .05$).

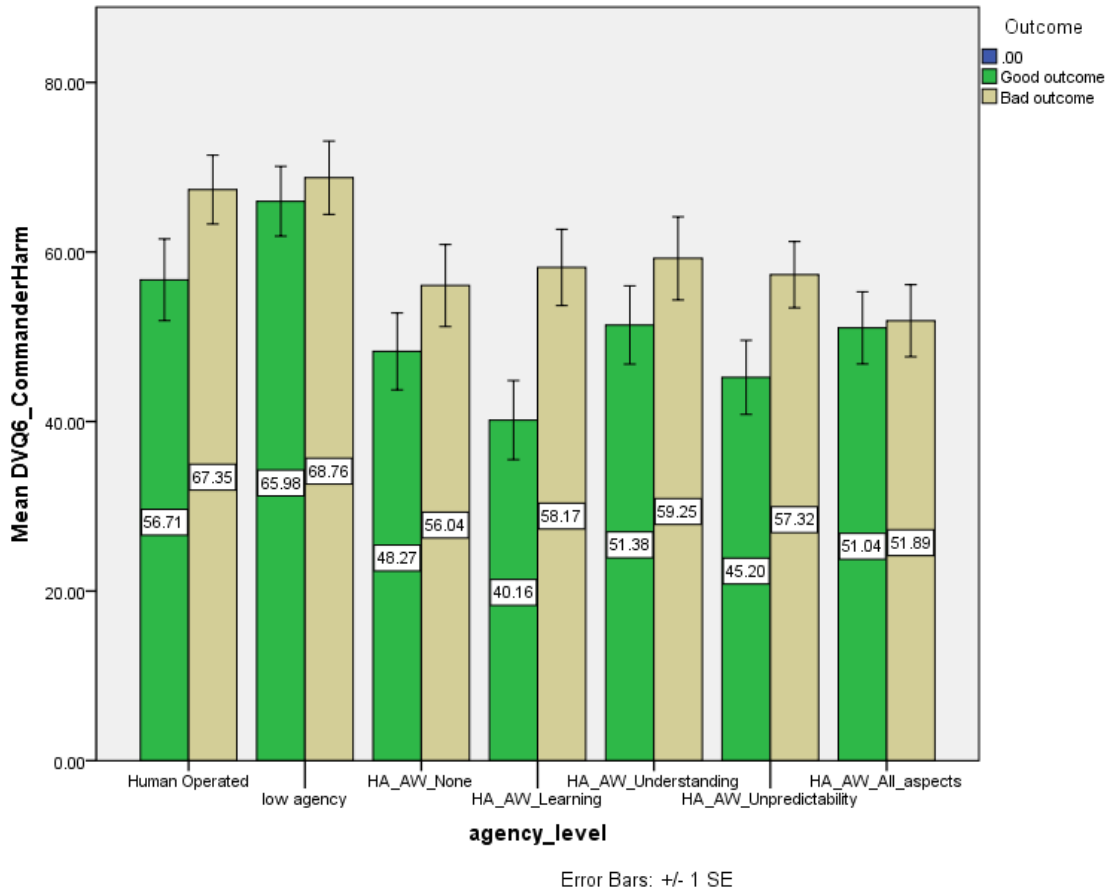


Figure 24 Mean value commander harm variable per condition per outcome

Human dignity

The Human Operator acts with much more respect for *human dignity* than the Autonomous Weapon except for the condition where the Autonomous Weapon is trained on a large data set for its mission and this difference is significant ($p < .05$). Also, it seems to matter when the Autonomous Weapon to make a mistake, because we observe significant ($p < .05$) differences between good and bad outcome scenarios for Autonomous Weapons for the low agency, high agency no aspects, and high agency all aspects scenarios.

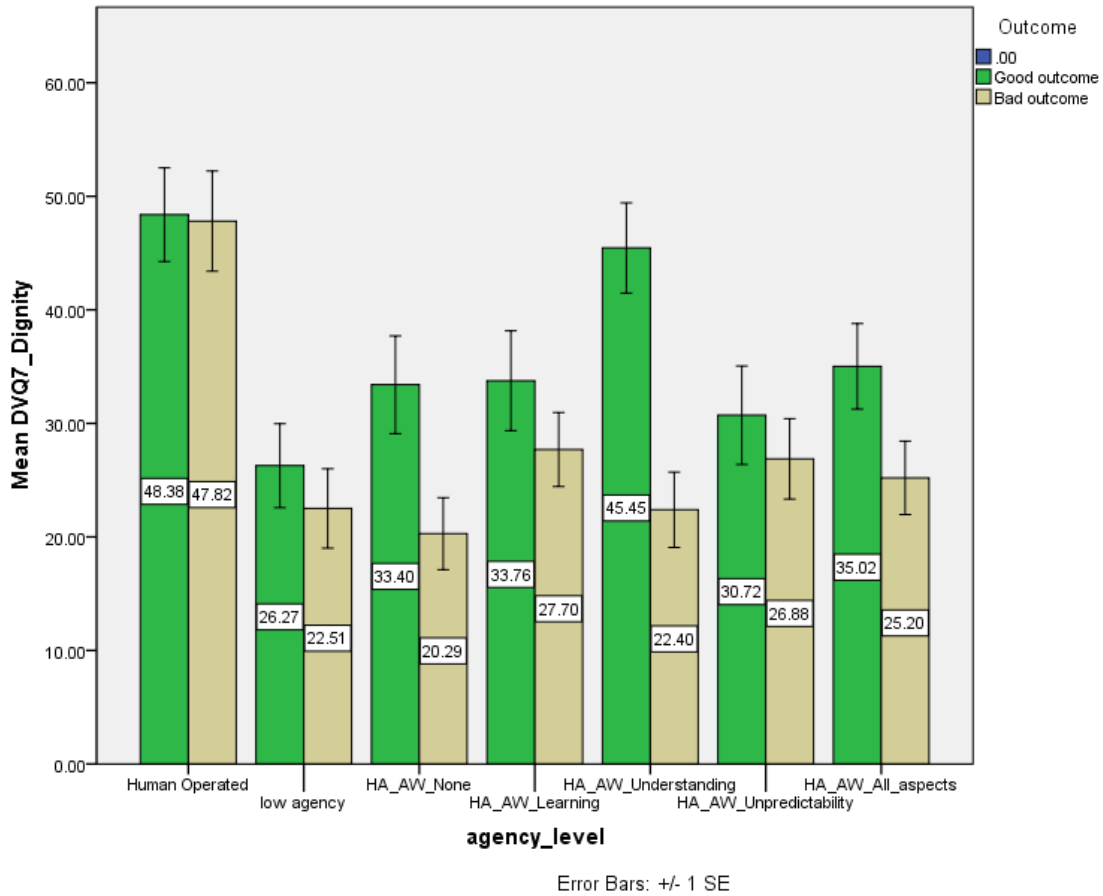


Figure 25 Mean value commander harm variable per condition per outcome

Confidence

Overall people have more *confidence* that the Autonomous Weapon will take the correct actions in the future after a good outcome than a bad outcome and all differences are significant ($p < .05$). They also have slightly more *confidence* in the Human Operator than in the Autonomous Weapon. The differences between these scenarios are significant ($p < .05$). When the Autonomous Weapon has learning capabilities, their confidence is a slightly higher compared to the other Autonomous Weapon conditions.

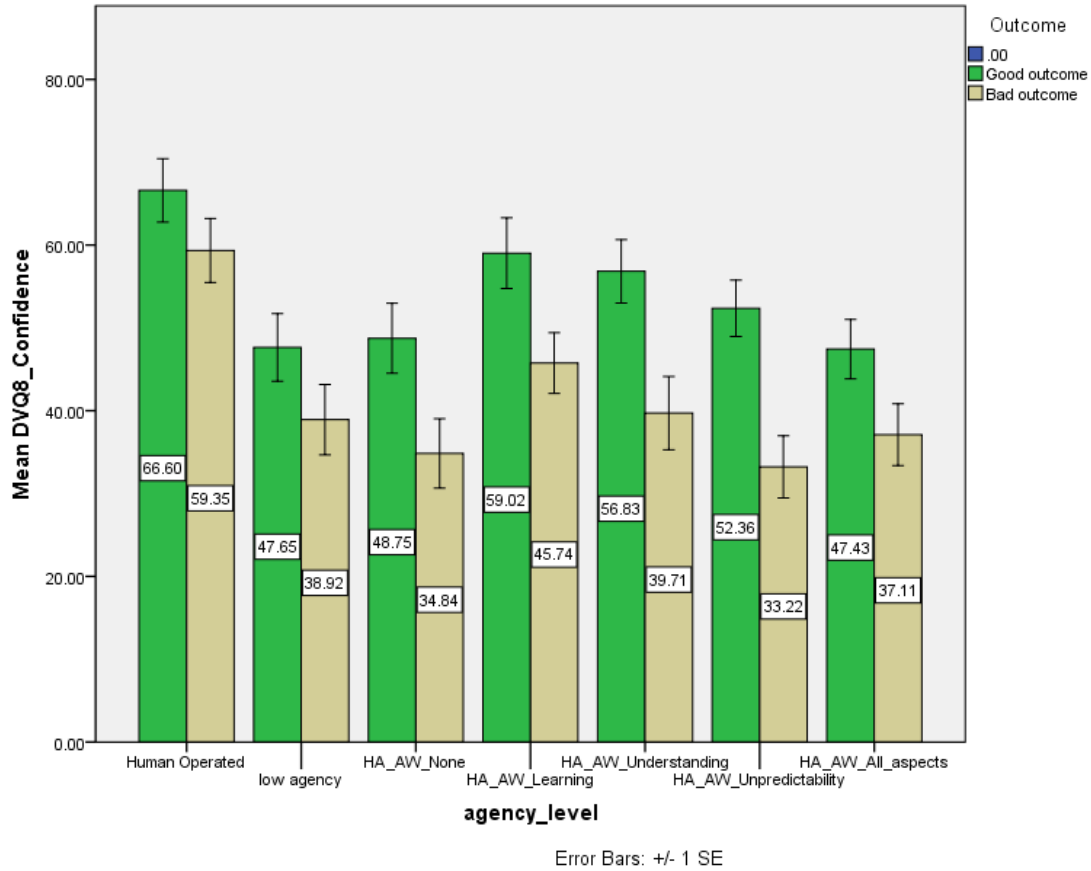


Figure 26 Mean value confidence variable per condition per outcome

Expectations

In good outcome scenarios both the Human Operator as the Autonomous Weapon perform more according to people's *expectations* than in bad outcome scenarios. All differences between the good outcome and bad outcome conditions are significant ($p < .05$). The expectations are lowest in the low agency condition when there is a bad outcome, but there is not much difference compared to the other conditions.

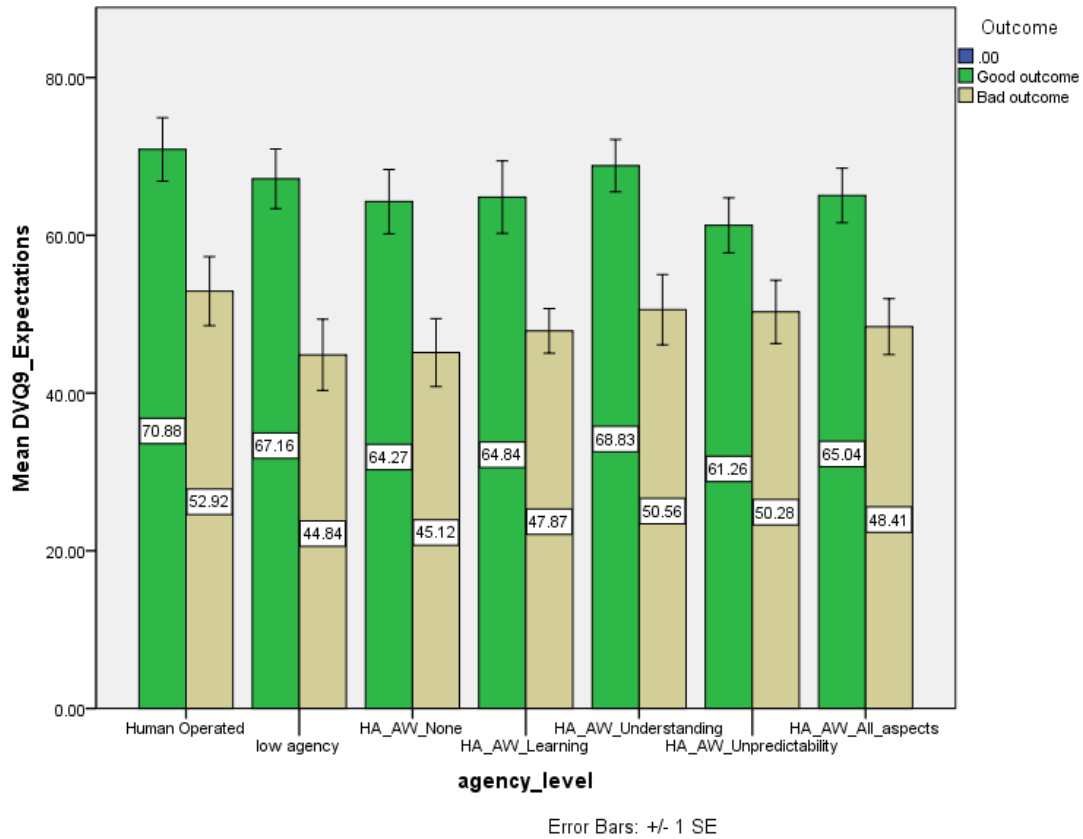


Figure 27 Mean value expectations variable per condition per outcome

Support

People have more *support* for Human Operated drones than for Autonomous Weapons. The *support* is lowest, but not significant, for the low agency conditions compared to the scenarios where information about the Autonomous Weapon is provided. Also, support for scenarios with a good outcome is higher than for bad outcome scenarios and except for the Human Operated and low agency Autonomous Weapon condition these differences are significant ($p < .05$).

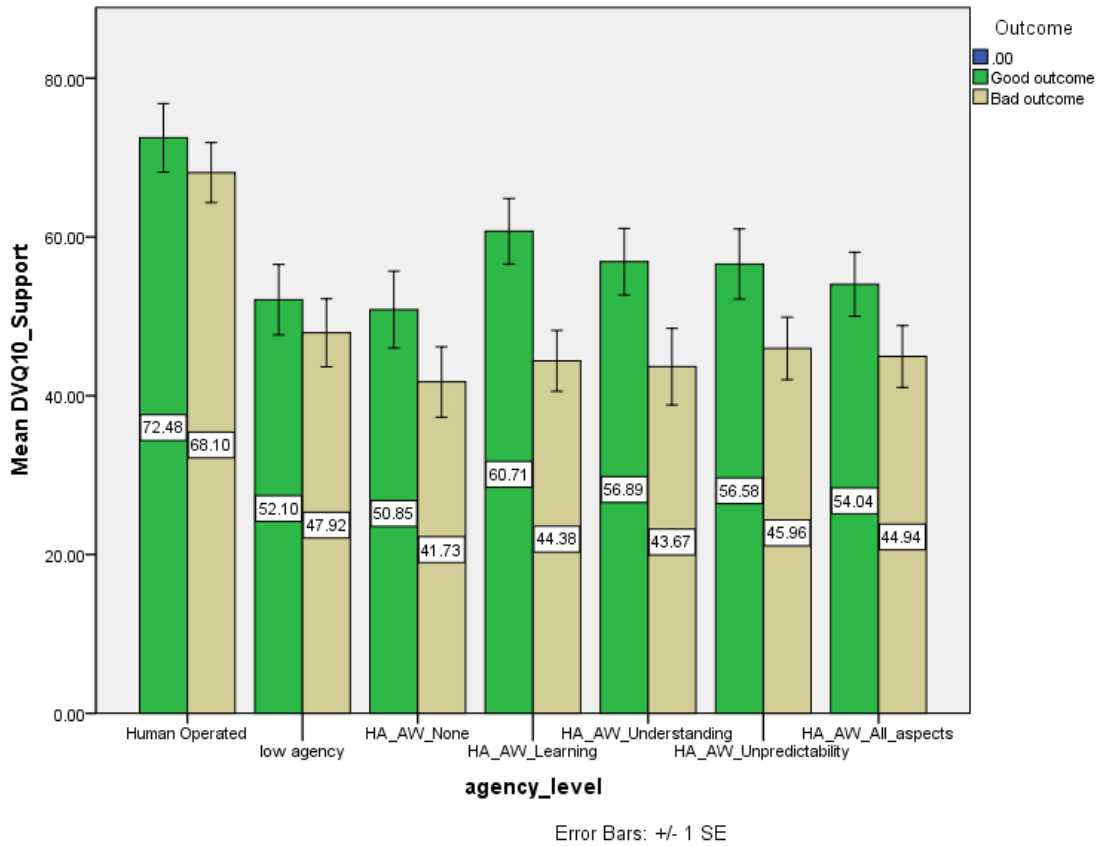


Figure 28 Mean value expectations variable per condition per outcome

Fairness

Overall, the actions of Human Operators are seen as more *fair* than those of Autonomous Weapons. The perception of *fairness* is much higher in good outcome scenarios compared to the bad outcome scenarios and in all scenarios these differences are significant ($p < .05$). The difference between the agency conditions of Autonomous Weapons is small (<10 points on a 100-point scale).

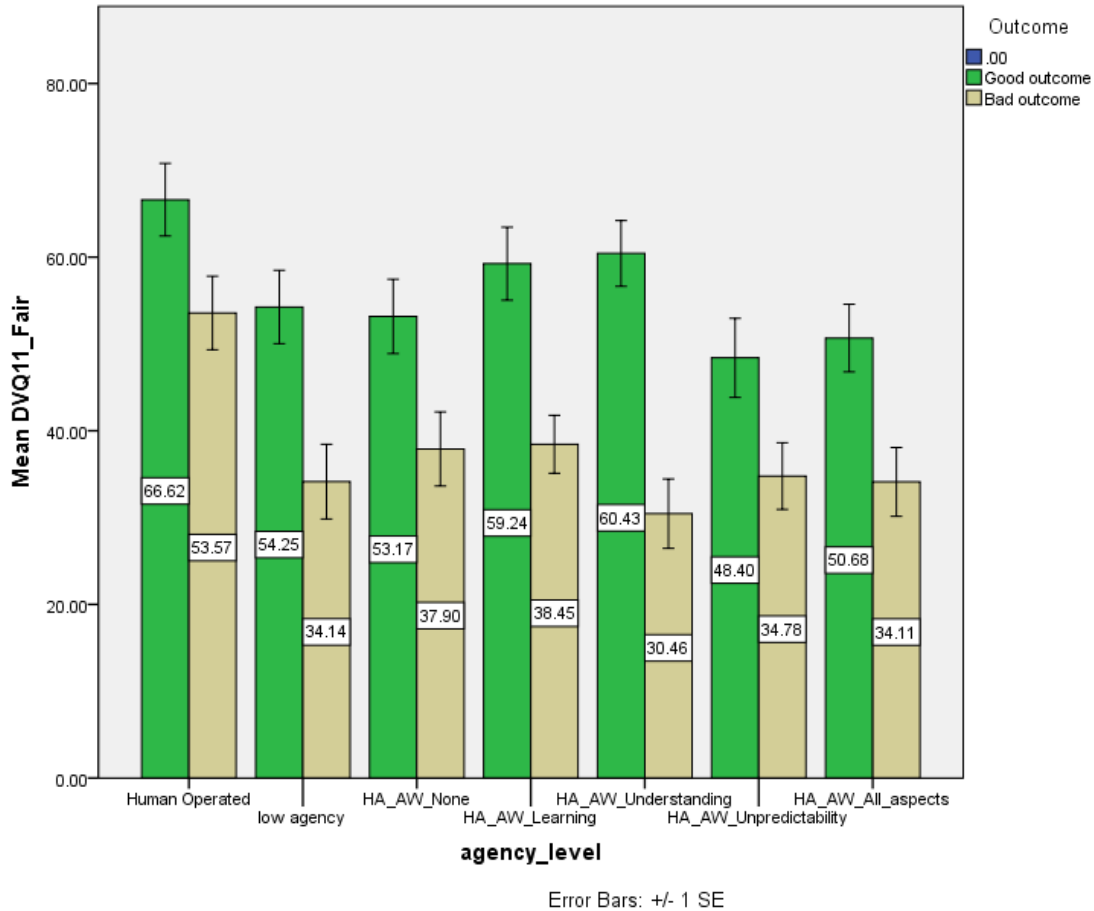


Figure 29 Mean value fairness variable per condition per outcome

Uneasy

In bad outcome scenarios people feel more *unease* with the actions of both the Human Operator as the Autonomous Weapon. There is not much difference between the level of anxiety for the different conditions of Autonomous Weapons. The anxiety level for the Human Operator is much lower than that for the Autonomous Weapon. Except for the high agency no extra aspects Autonomous Weapons condition, all differences between good and bad outcome scenarios are significant ($p < .05$).

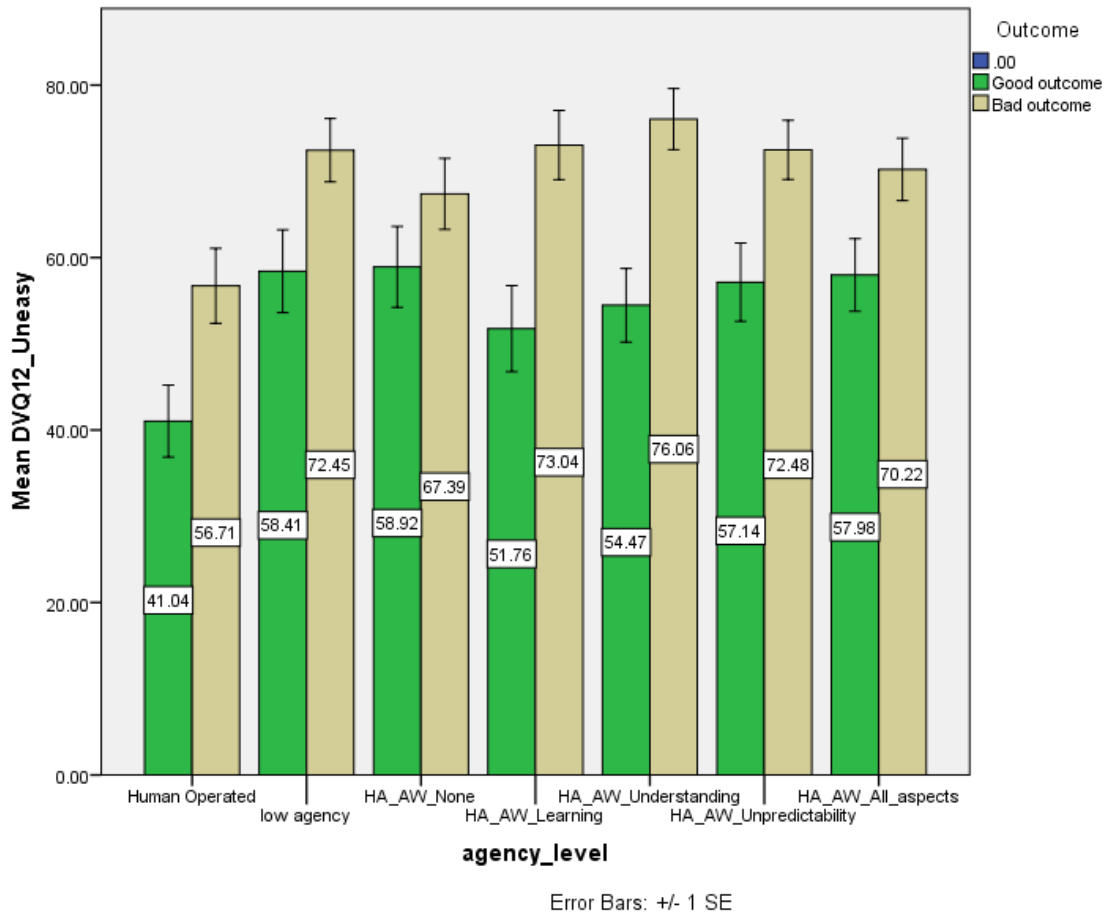


Figure 30 Mean value *unease* variable per condition per outcome

4.3.6. Conclusion pilot study 2

The results of pilot study 2 lead to the following conclusions:

- The reliability test showed that the four items of the agency construct are reliable and according to the PCA can be seen as one component;
- The correlation coefficient of the agency construct and all the dependent variables are significant and have the same directions as in pilot study one. We found a negative relationship between the agency construct and the level of *unease* meaning that the when the agency perception increases people indicate that they feel more uneasy, but this effect is still very small and approaching to 0 (r

= -.076). To see if we could get a bigger effect we changed the wording question of the *unease* variable in the final study;

- The agency manipulation shows that there is much difference between the high agency scenarios and the low agency scenario. The agency of the human operator and high agency Autonomous Weapons are at the same level. However, the mean of the agency construct for the high agency scenarios that include the algorithm aversion effects are almost the same. To draw out the algorithm aversion effect we need to do more literature and pilot studies. Due to time constraints, we did not pursue this line of research further and chose to focus the final study on the agency perception.
- Many of the effects that we found on the dependent variables were as we expected it to be based on the findings of pilot study 1. We did not observe distinct effects in the DV's between the different agency conditions of the Autonomous Weapon. We checked if we could replicate algorithm aversion effect that we found in pilot study 1 where *harm* was shifted from the drone to the *commander* in the low agency condition, but unfortunately this effect could not be observed in pilot study 2. Although this chain of responsibility needs to be further studied, we decided not to include it in the final study as we are not quite sure what is going on and more literature study and pilot studies are required.

4.4. Final study - military sample

To determine the number of scenarios for the final study, we performed Power calculations to estimate the total number of participants that we would need based on the results of the first pilot study, because we will use these scenarios also for the final study. Based on the Power calculations (effect size 0.4, a desired statistical power of 0.8 and a probability level of 0.05) we aimed for a total of 200 responses and determined that we could run 3 scenarios. We chose to focus on the agency perception and to compare the current technology, a Human Operated drone, to future technology, the high agency Autonomous Weapon (Table 15). We also added a neutral agency Autonomous Weapon scenario to compare the agency perception of the three conditions. In this study, we specifically asked about the drone and not the Human Operator controlling it in order to compare the agency perception of the artefacts. As we could only run three scenarios, we focused on the bad outcome condition, because the pilot studies showed that the effects most distinct in this condition.

We distributed the survey via email with an anonymous link and therefore had no guarantee on the number of respondents. After the data pre-processing and deleting the respondents who failed the attention check we were left 239 responses for the complete sample. The number of respondents per scenario ranged between 64 and 96 which is curious because the respondents should be distributed evenly over the scenarios. The sample of 239 is a combination of Dutch military and civilian personnel working at the MOD, because if we would exclude the civilians only 149 respondents are left which is too little for a robust data analysis.

Table 15 Scenarios final study

	<i>Agency levels</i>		
	Human Operated	Neutral agency AW	High Agency AW
Bad outcome	1	2	3

For this design, the following definitions are used:

- *Human operated drone*: drone directly controlled by a human. Hypothesized to be perceived as having very low agency.
- *High Agency AW*: participants are told that the Autonomous Weapon has agentic characteristics (such as weighing pros and cons and making independent choices). Hypothesized to be perceived as having very high agency.
- *Neutral Agency AW*: No information about agentic characteristics of the Autonomous Weapon are given. The strength of this approach is that we will be able to gather data on how participants use their pre-formed concepts of Autonomous Weapons to understand the actions of the Autonomous Weapon and compare those conceptions to the human operated and high agency Autonomous Weapon conditions.

4.4.1. Reliability analysis

For all four agency items $\alpha = .865$ which is above the threshold of 0.7 and cannot be improved by deleting an item which means that the agency construct is reliable (Table 16). Deletion of other items will decrease the α value and reduce the internal consistency.

Table 16 Results reliability analysis agency items

Cronbach's Alpha	N of Items
.865	4

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Thought	116.43	7175.573	.620	.863
Goal setting	123.33	6342.474	.756	.810
Act freely	122.62	6566.723	.740	.816
Achieve goals	118.84	6459.146	.740	.816

4.4.2. Principal component analysis (PCA)

The PCA shows that *Thought*, *Goal setting*, *Act freely* and *Achieve goals* can be viewed as one construct which accounts for 71.18 % variance in the original variables. This is also seen in the scree plot that shows the eigenvalues of the components (Figure 31). The PCA shows that the four items; *thought*, *goal setting*, *act freely* and *achieve goals*, can be viewed as one component of the agency construct (Table 17).

Table 17 PCA results agency items final study

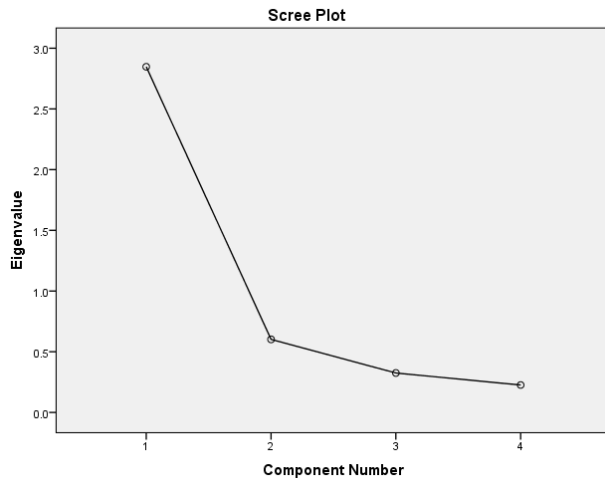


Figure 31 Scree plot PCA final study

Component Matrix^a

	Component	
	1	2
Thought	.774	.560
Goal setting	.873	
Act freely	.861	
Achieve goals	.863	

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

4.4.3. Correlation analysis agency construct

The correlation of the agency construct with the dependent variables is larger than in the previous pilot studies, but only significant for 7 out of 9 of the variables; *trust, human dignity, confidence, expectations, support, fairness* and *anxiety* ($p < .01$) (Table 18). There is a negative relationship between the agency construct and the level of uneasiness which is larger than previously in the pilot studies. It indicates that the when the agency perception increases people indicate that they feel more anxiety. The correlation analysis is not detailed enough to zoom in to these results, therefore this will be done in the analysis of the dependent variables.

4.4.4. Manipulation check on agency

The graph in Figure 32 shows an increase in agency perception over the conditions. The difference in the perception of agency between the Human Operated weapon scenario and the neutral agency Autonomous Weapon scenario, and the Human Operated scenario and the high agency Autonomous Weapon scenario is significant so these groups can be considered as different. The independent samples T-test for the neutral agency Autonomous Weapon scenario and the high agency Autonomous Weapon scenarios is with $p = .063$ just above the threshold for significance of $p < .05$. The independent samples T-test for the Human Operated condition and neutral agency Autonomous Weapon condition shows that $p = .001$ and for the Human Operated condition and the high agency condition Autonomous Weapon condition $p = .000$.

We hypothesized that the Human Operated drone will be perceived as having low agency and the high agency Autonomous Weapon will be perceived as having high agency and the results show that this is the case. The central analysis was concerned with how the neutral agency case compares with the other two cases. We hypothesized that military personnel will not perceive the neutral Autonomous Weapon as possessing mental states. Therefore, we expected no difference between the neutral agency Autonomous Weapon condition and the condition in which a human is operating a drone remotely. We also expected the neutral agency condition to be judged as significantly different from the high agency condition.

The results confirm that the Human Operated drone is perceived as having a low agency and that the high agency Autonomous Weapon is perceived as having high agency. However, the results also show that the

neutral agency Autonomous Weapon is perceived as having more agency than the Human Operated drone and that the agency perception in high agency condition is just slightly higher. This means that we will have to reject our hypothesis and conclude that military and civilians working at the Dutch MOD do perceive Autonomous Weapons as having agentic properties.

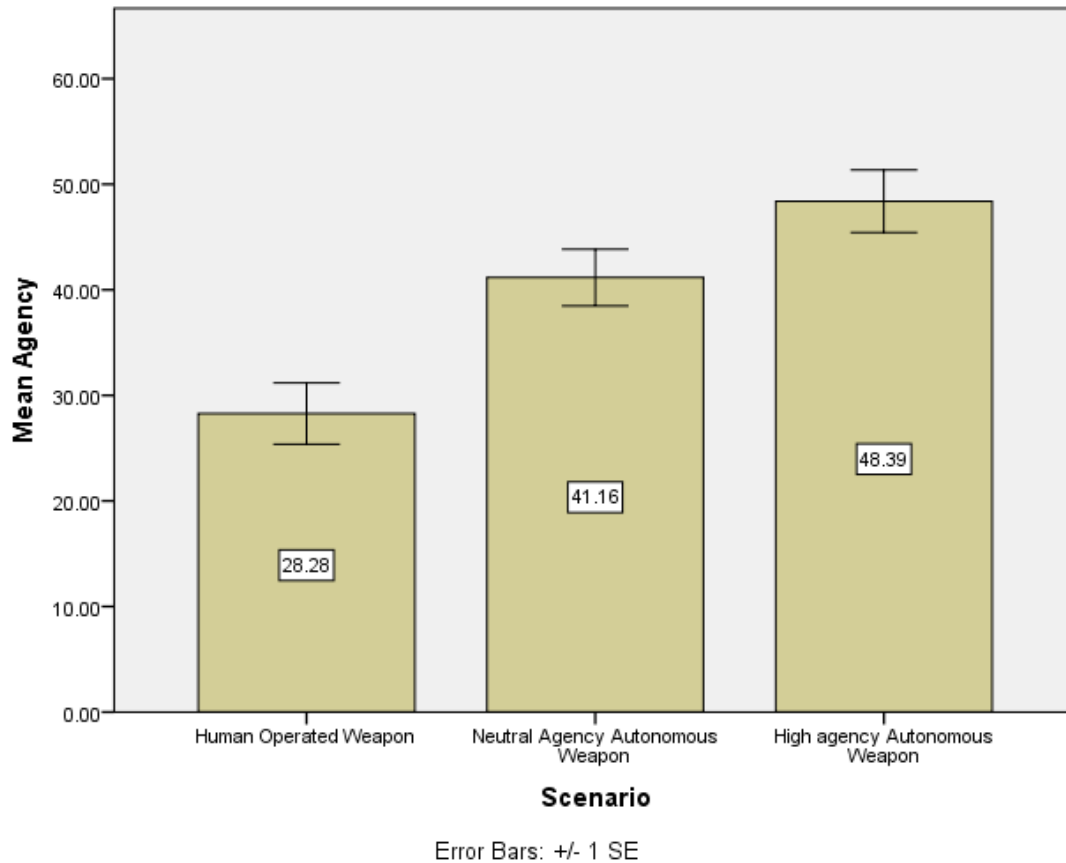


Figure 32 Mean value agency per condition

If we look at the difference in Figure 33 between the military and the civilian respondents we see something very interesting, but as the two groups are not significantly different in two of the scenarios, only in the high agency condition, we have to caution with interpreting the results so the following observation is illustrative. In both the Human Operated scenario as in the neutral agency scenario the agency perception of the military and civilian respondents is at the same level. However, in the high agency Autonomous Weapon scenario the military respondents perceive a lot more agency than their civilian counterparts. We looked at differences in demographics between the samples and these are remarkably similar, but one possible explanation could be that the civilian sample has 10% more respondents that worked with AI than the military sample. However, we are not sure that this explains the difference in agency perception between the two respondent groups and this aspect will need further investigation.

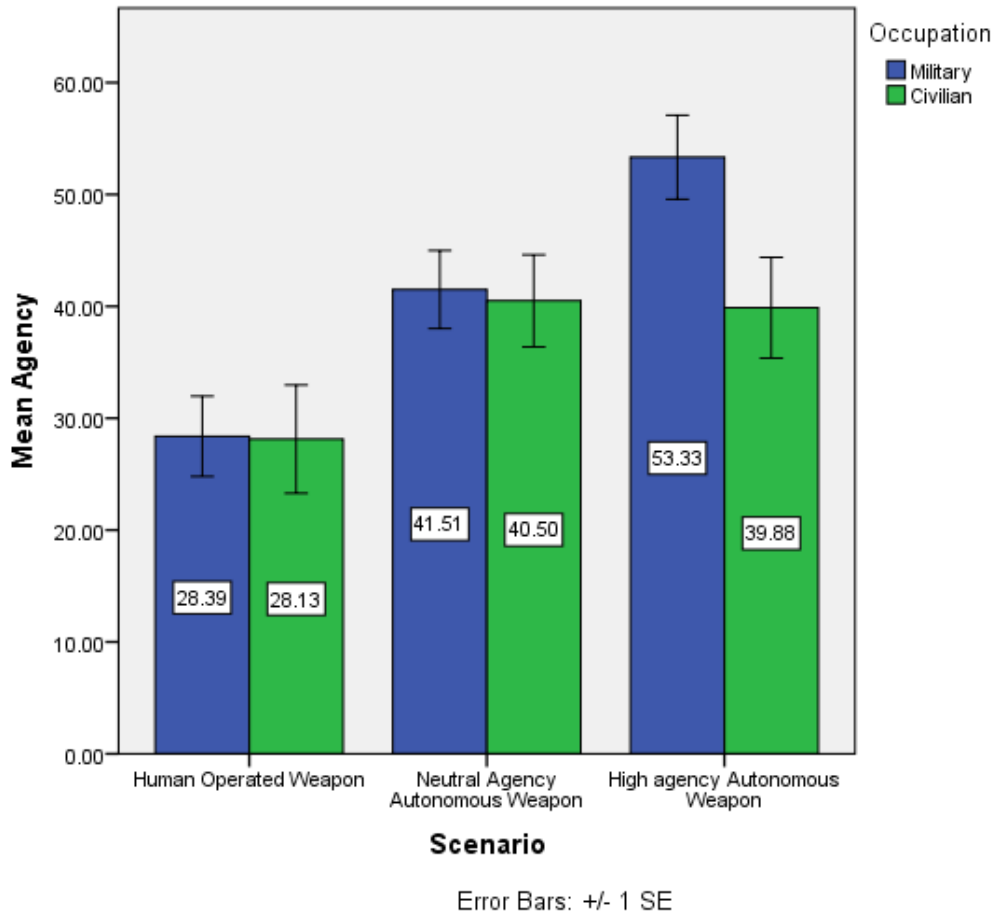


Figure 33 Mean value agency per condition per group

Table 18 Correlations agency construct with dependent variables for final study

		Correlations									
		Agency	DVQ1_Blame	DVQ2_Trust	DVQ3_Harm	DVQ4_Human_Dignity	DVQ5_Confidence	DVQ6_Expectations	DVQ7_Support	DVQ8_Fair	DVQ9_Anxiety
Agency	Pearson Correlation	1	.009	.406**	-.070	.347**	.428**	.320**	.337**	.451**	-.216**
	Sig. (2-tailed)		.888	.000	.275	.000	.000	.000	.000	.000	.001
	N	248	248	248	248	248	248	248	248	248	248
DVQ1_Blame	Pearson Correlation	.009	1	-.039	.056	.018	.027	-.109	-.002	-.080	.128*
	Sig. (2-tailed)	.888		.545	.379	.783	.667	.086	.979	.209	.044
	N	248	248	248	248	248	248	248	248	248	248
DVQ2_Trust	Pearson Correlation	.406**	-.039	1	-.217**	.482**	.767**	.324**	.597**	.540**	-.451**
	Sig. (2-tailed)	.000	.545		.001	.000	.000	.000	.000	.000	.000
	N	248	248	248	248	248	248	248	248	248	248
DVQ3_Harm	Pearson Correlation	-.070	.056	-.217**	1	-.180**	-.197**	-.155*	-.133*	-.200**	.360**
	Sig. (2-tailed)	.275	.379	.001		.004	.002	.015	.036	.002	.000
	N	248	248	248	248	248	248	248	248	248	248
DVQ4_Human_Dignity	Pearson Correlation	.347**	.018	.482**	-.180**	1	.564**	.239**	.564**	.472**	-.385**
	Sig. (2-tailed)	.000	.783	.000	.004		.000	.000	.000	.000	.000
	N	248	248	248	248	248	248	248	248	248	248
DVQ5_Confidence	Pearson Correlation	.428**	.027	.767**	-.197**	.564**	1	.376**	.685**	.600**	-.495**
	Sig. (2-tailed)	.000	.667	.000	.002	.000		.000	.000	.000	.000
	N	248	248	248	248	248	248	248	248	248	248
DVQ6_Expectations	Pearson Correlation	.320**	-.109	.324**	-.155*	.239**	.376**	1	.271**	.475**	-.264**
	Sig. (2-tailed)	.000	.086	.000	.015	.000	.000		.000	.000	.000
	N	248	248	248	248	248	248	248	248	248	248
DVQ7_Support	Pearson Correlation	.337**	-.002	.597**	-.133*	.564**	.685**	.271**	1	.528**	-.488**
	Sig. (2-tailed)	.000	.979	.000	.036	.000	.000	.000		.000	.000
	N	248	248	248	248	248	248	248	248	248	248
DVQ8_Fair	Pearson Correlation	.451**	-.080	.540**	-.200**	.472**	.600**	.475**	.528**	1	-.440**
	Sig. (2-tailed)	.000	.209	.000	.002	.000	.000	.000	.000		.000
	N	248	248	248	248	248	248	248	248	248	248
DVQ9_Anxiety	Pearson Correlation	-.216**	.128*	-.451**	.360**	-.385**	-.495**	-.264**	-.488**	-.440**	1
	Sig. (2-tailed)	.001	.044	.000	.000	.000	.000	.000	.000	.000	
	N	248	248	248	248	248	248	248	248	248	248

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

4.4.5. Dependent variables analysis

The correlation between the agency construct and the dependent variables *trust*, *human dignity*, *confidence*, *expectations*, *support*, *fairness* and *anxiety* are significant ($p < .01$) therefore the most striking findings for each of these 7 dependent variables are described in this section.

Trust

Overall people have more *trust* that Human Operators will take correct actions in the future than Autonomous Weapons and the difference between these groups is significant ($p < .05$). The level of *trust* for the neural agency and high agency Autonomous Weapon condition is equal and the difference between these groups is not significant.

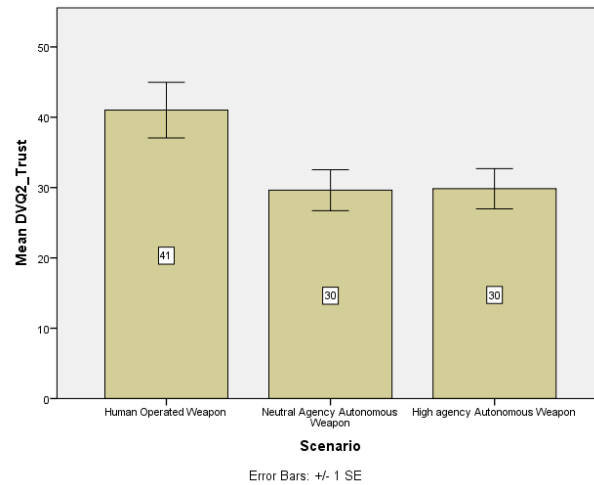


Figure 34 Mean value trust variable per condition

Human dignity

Human Operated weapons are perceived to act with more respect for *human dignity* than Autonomous Weapons and the difference between these groups is significant ($p < .05$). The difference between the neutral and high agency conditions for the Autonomous Weapons is small and the difference between these groups is not significant.

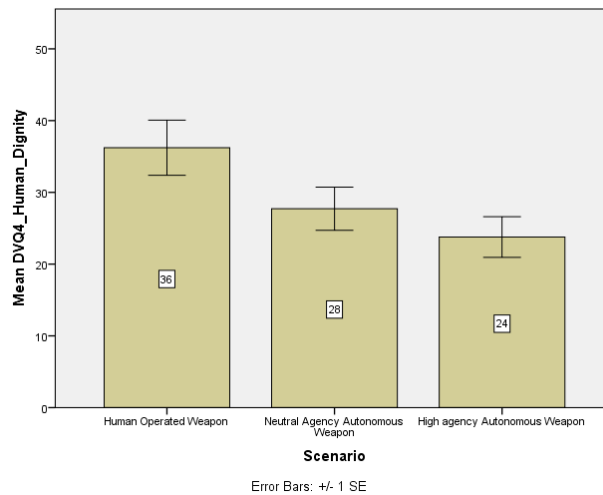


Figure 35 Mean value human dignity variable per condition

Confidence

People have more *confidence* in that Human Operated drones will take correct actions in the future than in Autonomous Weapons and the difference between these groups is significant ($p < .05$). There is no significant difference between the low and high agency conditions for the Autonomous Weapons.

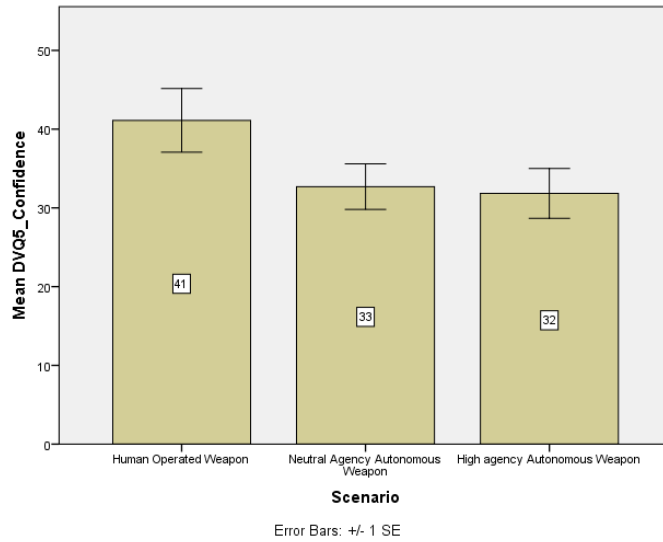


Figure 36 Mean value confidence variable per condition

Expectations

A slightly lower difference can be observed in the *expectations* of people between the Human Operated condition and both the Autonomous Weapons conditions. The level of expectations of the Autonomous Weapon conditions is equal. However, the difference between all groups is not significant.

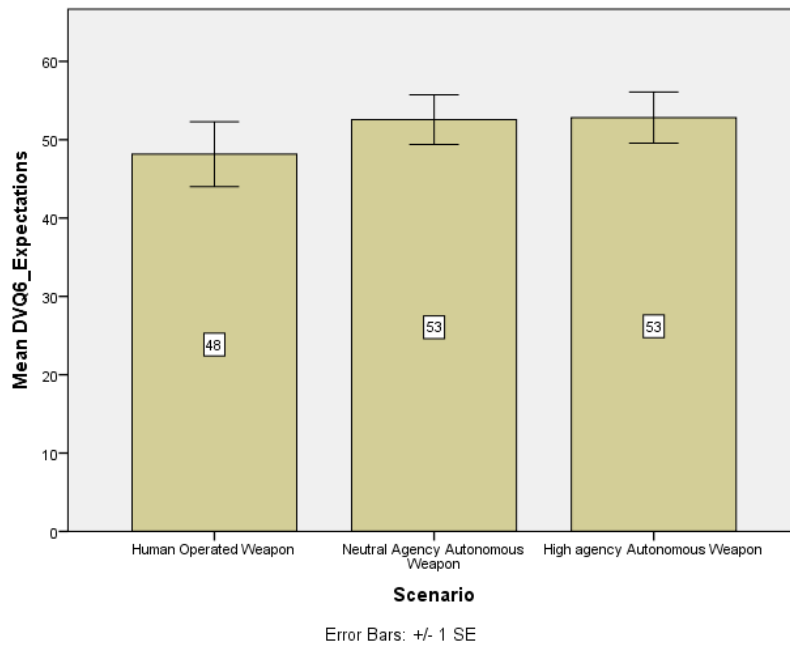


Figure 37 Mean value expectations variable per condition

Support

In general, Human Operated drones are more *supported* than Autonomous Weapons and the difference between these groups is significant ($p < .05$). For Autonomous Weapons, it seems that the agency perception does not influence the level of *support* and the difference between the neutral and high agency group is not significant.

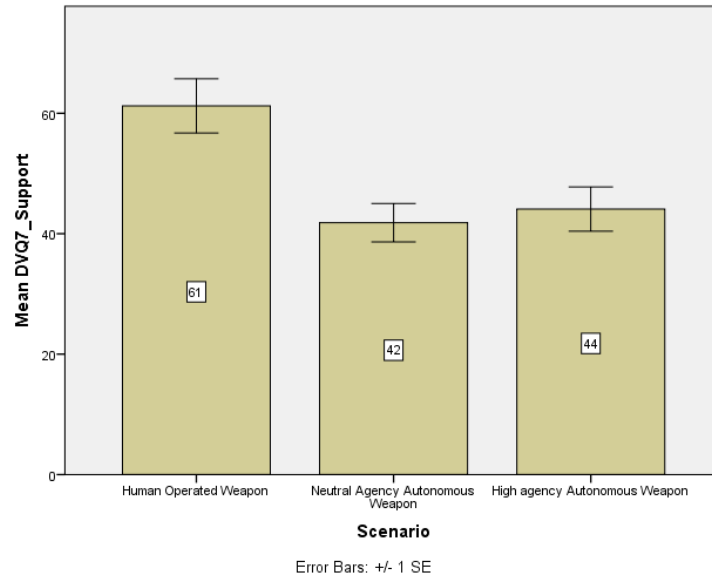


Figure 38 Mean value support variable per conditions

Fairness

The actions of both the Human Operated drone as the Autonomous Weapons are considered to be equally *fair* which is a striking result because we expected that the actions of the Human Operated drone might be viewed as more *fair* than that of an Autonomous Weapon. Unfortunately, the difference between these groups is not significant.

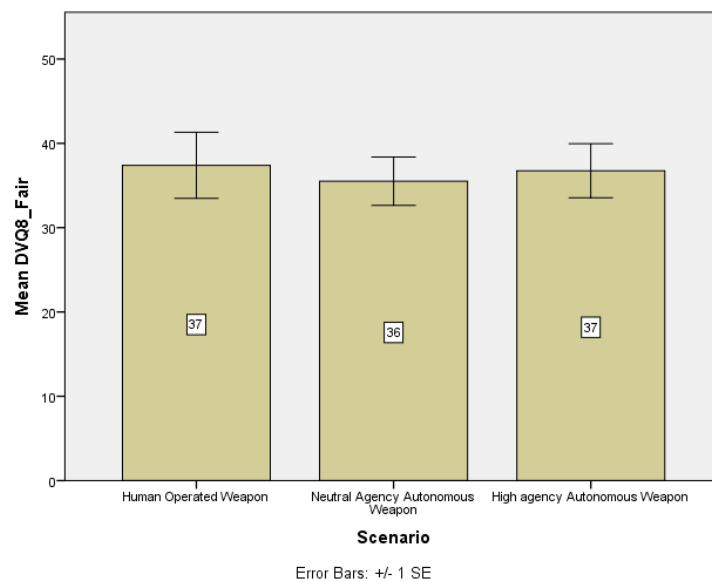


Figure 39 Mean value fairness variable per condition

Anxiety

People are more *anxious* about Autonomous Weapons than a Human Operated weapon and the difference between these groups is significant ($p < .05$). The neutral agency and higher agency Autonomous Weapon scenarios show a small increase in the level of anxiety, but the difference between these groups is not significant.

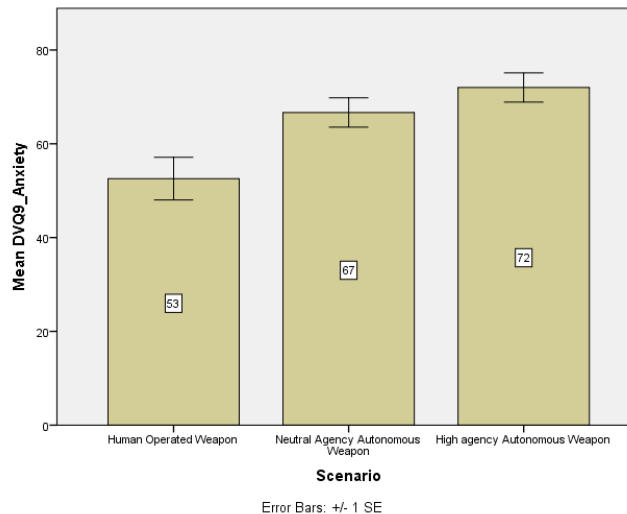


Figure 40 Mean value anxiety variable per condition

4.4.6. Conclusions final study

The results of the final study lead to the following conclusions:

- The reliability test showed that the four items of the agency construct are reliable and according to the PCA can be seen as one component;
- The agency perception of the neutral agency Autonomous Weapons condition is higher than the agency perception of the Human Operated drone condition. The agency perception in the high agency condition is higher than the neutral agency condition, but the distinction between the groups is not significant with a $p = .063$ (Figure 32);
- In both the Human Operated scenario as in the neutral agency scenario the agency perception of the military and civilian respondents is at the same level. However, in the high agency Autonomous Weapon scenario the military respondents perceive a lot more agency than their civilian counterparts (Figure 33) but as the two groups are not significantly different in two of the scenarios, only in the high agency condition, we have to caution with interpreting the results and in drawing conclusions. We are not sure what causes this the difference in agency perception between the two respondent groups and this aspect will need further investigation.
- The correlation analysis shows that 7 of the 9 dependent variables are significantly correlated to the agency construct. These 7 are *trust*, *human dignity*, *confidence*, *expectations*, *support*, *fairness* and *anxiety* (Table 18);
- Based on the more detailed analysis of the dependent variables, we can observe the level of *trust*, *confidence*, *human dignity* and *support* in the actions taken by Human Operated drones is higher than those taken by Autonomous Weapons (Figure 34, Figure 35, Figure 36, Figure 38).
- The level of expectations and fairness for Human Operated drones and Autonomous Weapons are equal (Figure 37, Figure 39).
- People have higher levels of anxiety for Autonomous Weapons than for Human Operated weapons and these levels are highest in the high agency condition.

5. Design of Moral Machine for Autonomous Weapons

I've watched Eye In The Sky..this survey reminds me of that movie.

Respondent pilot study 1

In the previous section we described the scenarios that we tested in the pilot and final studies. Although the data samples were large enough to perform statistical analyses, the studies had only 50 to 96 respondents per scenario and these respondents had very specific demographical characteristics. For example, the respondents on MTurk are primarily in the age range 25 – 35 with a higher college level education and the military study consisted for 95% of man with a Dutch nationality. To be able to generalize the results, the study needs more respondents that represent a larger demographic group. We propose a design for a Moral Machine for Autonomous Weapons for a large scale study of this topic for which we build on the concept of the Moral Machine, that was developed by the Scalable Cooperation group of the Media Lab at MIT (Scalable Cooperation Group, 2016).

The proposed design of the Moral Machine for Autonomous Weapons is part of the technical investigation phase of the Value-Sensitive Design method. We utilise this phase to design technology to take the next step in our research and build on the scenarios used in the pilot and final studies to gain insight in the moral judgement of people regarding to Autonomous Weapons. Creating a website that hosts the Moral Machine for Autonomous weapons would allow us to collect data on moral judgements on a large scale from diverse demographic groups which could be used for more robust and generalizable results. Large scale data collection in multiple countries might reveal cultural differences in moral judgement of Autonomous Weapons. For example, the views in Western countries on the deployment of these type of weapons is most likely different than the views in Middle-Eastern or Asian countries. The data from these different countries can be used in the debate of Autonomous Weapons which is in our opinion currently dominated by Western viewpoints.

This section first describes the Moral Machine for Autonomous Vehicles, followed by our proposal for the Moral Machine for Autonomous Weapons for which we specify the scenarios and variables and to conclude we provide some pointers for the implementation the website.

5.1. Moral machine for Autonomous Vehicles

The original Moral Machine is a '*... platform for gathering data on human perception of the moral acceptability of decisions made by autonomous vehicles faced with choosing which humans to harm and which to save.*' (Awad, 2017, pp. 42-43). The website has three modes for users: 1) a *Judge* mode in which users can decide the outcome for 13 series of scenarios, 2) a *Design* mode in which users can design their own scenarios, and 3) a *Browse* mode in which users can view the scenarios of others. The main feature is the *Judge* mode in which users choose between two scenarios that contain different variables of: *Characters* {gender, social value, age, species, fitness, utilitarianism}, *Interventionism* {omission, commission}, *Relationship to vehicle* {driver, pedestrian} and *Concern for law* {legal action, illegal action}. Other features include a video describing the project, instructions and background information for the users.

The Moral Machine is set-up as a Massive Online Experiment (MOE) (Reips, 2002) which is aimed at recruiting a large sample pool with a diverse background in a short amount of time at a low cost. These

features are clear advantages of a MOE, but the downside is that the conditions are hard to control, as users can take surveys multiple times and are self-selected, which means that they can join the experiment and drop out whenever they want (Awad, 2017). The Moral Machine was developed by means of a rapid-prototyping method on the Meteor platform which offers dynamic scripting, template-based structures and has a high responsiveness. It is hosted on a cloud application service, is intended for usage on mobile devices and is optimized for social media sharing with Cards, Markup and Open Graph tags. By May 2017, approximately 3 million users over 160 countries assessed over 30 million scenarios making it one of the biggest large-scale moral judgement tools that exist.

5.2. Moral Machine for Autonomous Weapons

Autonomous Weapons are a sensitive topic and we observed that it invokes a primary response of anxiety and unease with people. In our opinion, it would not be prudent to develop an open platform to gather data on a large-scale at first, because an open platform could attract negative sentiment and unwanted actions which would be counterproductive for our research, for example people creating scenarios in which certain groups of people, such as Muslims or Women, are specifically targeted. Therefore, it is advisable to take a more step-wise approach in scaling up to a Massive Online Experiment. To run a large-scale follow-up study of the moral judgements of people regarding Autonomous Weapons a controlled experiment with a limited set of conditions and sample needs to be designed. These conditions can be visualized in scenarios that allow users to take the survey after obtaining a password via a web-interface to a secure server. After gathering initial data and user feedback, the next step could be to scale up to a large-scale open platform, like the Moral Machine, where people can judge the scenarios to collect a large amount of data in several countries. However, due to the sensitivity of the topic we believe would not be advisable to allow people to create their own scenarios or share their results on social media as the *Design* feature of the original Moral Machine offers.

We propose the following features for the design of the Moral Machine for Autonomous Weapons:

- A *Judge* mode where users can choose between different scenarios and indicate which of the two scenarios is most morally acceptable;
- Conditions need to differ on only one variable at a time in the comparison of scenarios so that the measurement cannot be attributed to the result of several conditions to prevent confounding effects;
- A page with a brief overview and explanation of the meaning of the variables before the user starts judging the scenarios;
- A page with more information on the specific scenario that the user is judging which can be accessed if he or she wants more explanation for clarity;
- After judging the scenarios, the results are presented to the user;
- A visualisation of the user's results compared to the results of other users to provide the user feedback on their judgement;
- A second round of judging to allow users to judge the scenarios again so that they can alter their moral judgement and views based on the comparison of their results to others.

5.3. Scenarios

In this section, the variables for the Moral Machine for Autonomous Weapons are defined which are depicted in two example scenarios. The variables are based on the scenarios of the pilot and final studies which describe a military convoy that is supported by a drone in the air and they are inspired by the variables used in the Moral Machine for Autonomous Vehicles.

5.3.1. Variables for scenarios

For the prototype of the Moral Machine for Autonomous Weapons 6 variables are derived from either the pilot and final studies, for example *Type of Weapon* and *Outcome*, or inspired by the Moral Machine for Autonomous Vehicles, like *Character* and *Number of Characters*. The 6 variables are: *Type of Weapon (W)*, *Location (L)*, *Character (C)*, *Number of Characters (N)*, *Outcome (O)* and *Mission (M)*. The variables described in the next paragraphs are examples and suggestions and will need to be professionally designed when implemented in a Moral Machine for Autonomous Weapons. For example, the *Location* variable is based on a cartoon and game like depiction of a desert and village and it needs to be tested if this is a clear representation or if a photograph is more suitable when used in a study.

Type of Weapon

This dimension shows the Type of Weapon (W) that is deployed to support the convoy. This dimension tests if people judge the current technology, the human operated drone, to be more morally acceptable than an Autonomous drone, which is future technology. This can be used to get insight in the support for the technologies. The Type of Weapon is either a Human Operated drone (left hand side) or an Autonomous Weapon (right hand side) (Figure 41). If *Type of Weapon* is the discriminative variable then the different pictograms are shown on the scenario for people to choose between them as is shown in example 1 (Figure 47). Otherwise the same Type of Weapon is depicted on the scenario of example 2 (Figure 48).

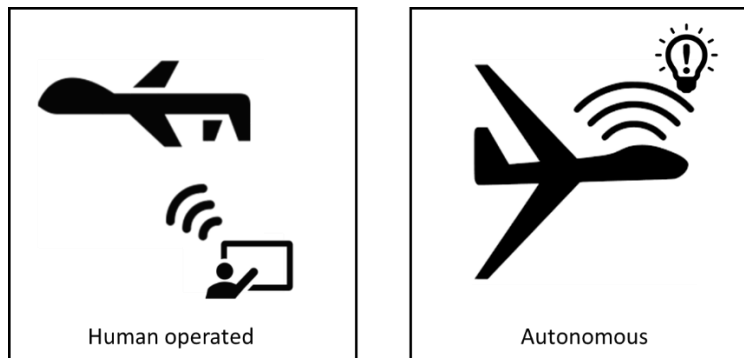


Figure 41 Type of weapon variable $W = \{\text{human operated drone, autonomous drone}\}$

Location

This dimension shows the Location (L) as setting for the scenario which can either be in the desert (on the left) or in the village (on the right). In this dimension, we test which location is more morally acceptable for people to deploy the Autonomous Weapon or Human Operated drone. If *Location* is the discriminative variable then both different images are shown on each of the scenario as is shown in example 2 (Figure 48). Otherwise the same location is depicted in both scenarios as in example 1 (Figure 47).



Desert



Village

Figure 42 Location variable $L = \{desert, village\}$

Character

This dimension shows the type of Character (C) that is involved as bystanders in the scenario which can either be a man (on the left), a woman (in the middle) or a child (on the right) (Figure 43). By varying the characters, we gain insight which bystanders people find morally acceptable when an Autonomous or Human Operated weapon is used. If *Character* is the discriminative variable then different characters are shown on each of the scenario. Otherwise the same characters are depicted in both scenarios.

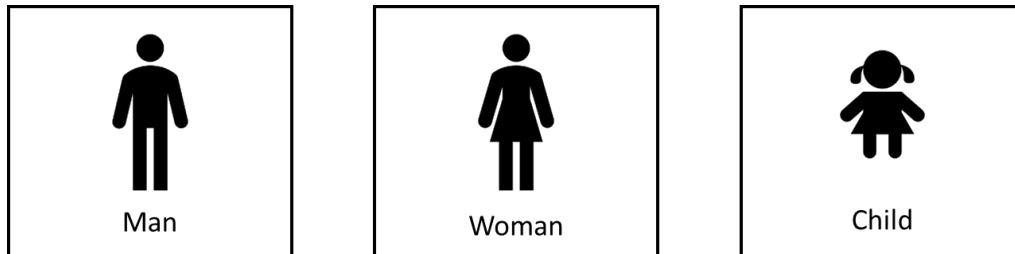


Figure 43 Character variable $C = \{man, woman, child\}$

Number of Characters

This dimension shows the *Number of Characters* (N) that are involved in the scenario which can range from one character (on the left) up to five characters (on the right) (Figure 44). This dimension allows us to gain insight into how many bystanders people find morally acceptable when an Autonomous or Human Operated Weapon is used. If the *Number of Characters* is the discriminative variable then different numbers of characters are shown on each of the scenario. Otherwise the same number of characters is depicted in both scenarios.

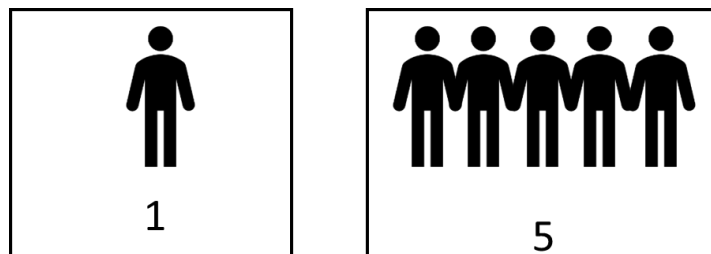


Figure 44 Number of Characters variable $N = \{1..5\}$

Outcome

This dimension shows the Outcome (O) of the scenario which can either be no collateral damage (on the left) or an outcome with collateral damage of the number of characters involved in the scenario (on the right) (Figure 45). This allows us to gain insight into how the outcome influences the moral acceptability of when an Autonomous or Human Operated Weapon is deployed. If *Outcome* is the discriminative variable then different pictograms are shown on each of the scenario. Otherwise the same outcome variable is depicted in both scenarios.

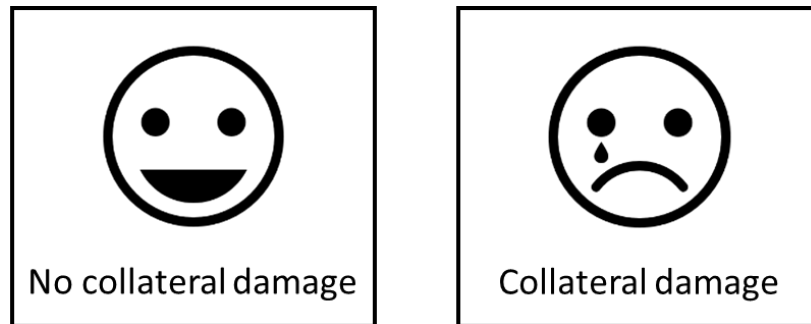


Figure 45 Outcome variable $O = \{\text{collateral damage, no collateral damage}\}$

Mission

This dimension shows the Mission (M) of the weapon in the scenario which can either be to defend the convoy only when a direct threat is perceived (on the left) or to attack vehicles that are on the target list even as they do not pose a direct threat for the convoy (on the right) (Figure 46). In this dimension, we test which type of mission is morally acceptable to people. If *Mission* is the discriminative variable then different pictograms are shown on each of the scenario. Otherwise the same mission variable is depicted in both scenarios.

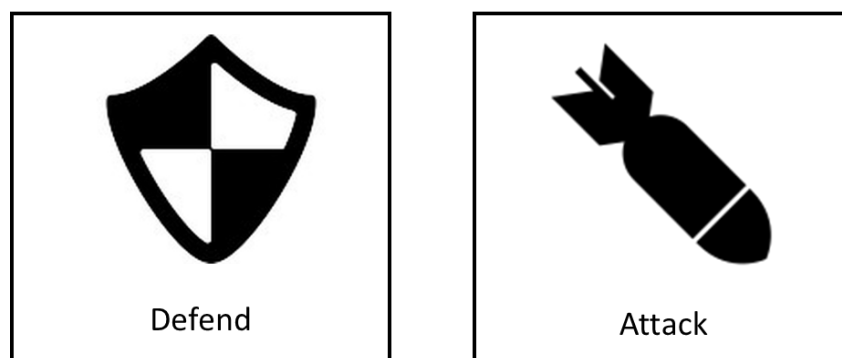


Figure 46 Mission variable $M = \{\text{defend, attack}\}$

5.3.2. Example Scenarios

The variables described above can be used to create scenarios in which each scenario differs on only one variable. The question posed in the scenario is the same question as that is being asked when judging the scenarios in the original Moral Machine. In this section, we depict two scenarios as example to show the

concept, but for the sake of brevity chose not to show all variables in an endless list of examples. Example 1 can be seen in Figure 47 which shows a convoy in a desert location. The difference between the scenarios is that on the left the convoy is defended by a Human Operated drone and in the scenario on the right by an Autonomous Weapon. In both cases there is one person near the road which is killed by collateral damage.

Which scenario is most acceptable to you?

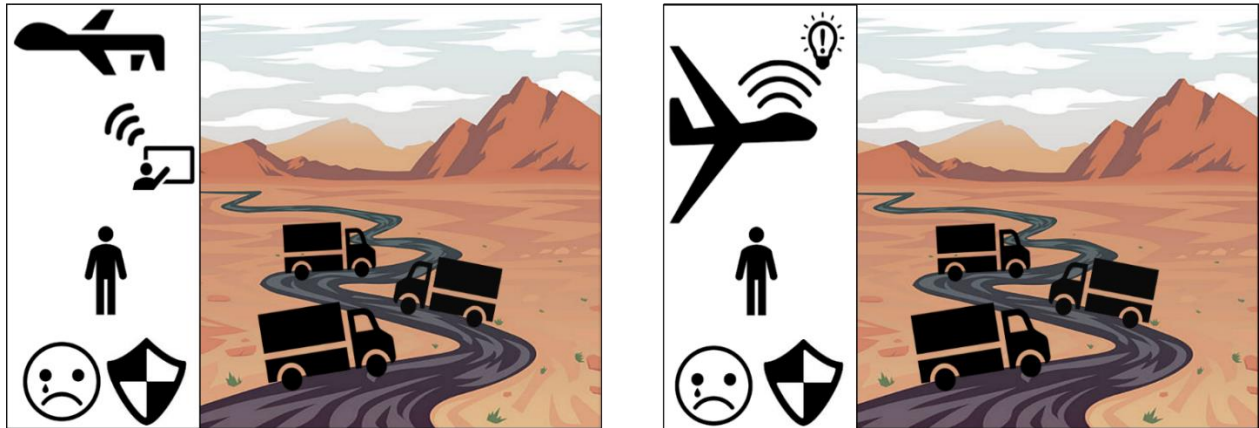


Figure 47 Example scenario 1

In the next example scenario (Figure 48) both convoys are defended by an Autonomous Weapon, but the location is different. The convoy on the left drives through a desert and the one on the right through a village. In both situations two children are playing near the road and the attack causes no collateral damage.

Which scenario is most acceptable to you?

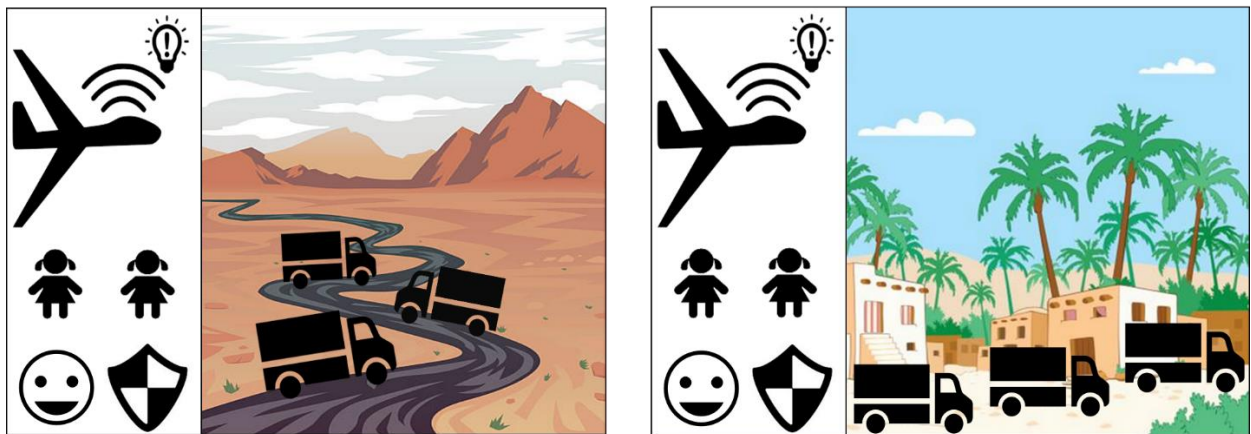


Figure 48 Example scenario 2

5.4. Implementation

The implementation of the Moral Machine for Autonomous Weapons requires several activities and is a project that will take at least six months to complete and another year to run the study. We briefly describe the phases in this section.

In phase one, the scenarios and variables sketched in the previous sections need to be designed in which attention is paid to the clarity and interpretability of the variables. According to the researcher who build the original Moral Machine the process of creating the figures for the scenarios took approximately three months and was done in several iterations by a professional graphic designer.

In phase two, the infrastructure for the website needs to be developed and build which will also take several months. Given the sensitivity of the topic, a secure website is needed to run the scenarios so that it can only be accessed with a password. This requires a mechanism to distribute the passwords to people who interested taking the survey. For example, a website that sends a link to a survey after registration. As we intend to upscale to a Massive Online Experiment, this possibility should be taken into account as a requirement from the start of the project. This means that the website should be hosted in a server environment that allows for dynamic upscaling and down scaling based on the number of active users. At the same time, it is important that the security of the website and server is strong and that the research data is owned by the University that runs the study. These requirements imply that the website should not be hosted by a commercial corporation, such as Microsoft, Google or Amazon who offer such dynamic cloud services, but on server owned by the University.

The third phase is that these scenarios need to be tested in several pilot studies to check whether they generate useful results and will have to be adjusted if it turns out that this is not the case. This will be a process that will take several iterations until the final study can be tested. The original Moral Machine collected data from June 23, 2016 until May 2017 (Awad, 2017) and it is advised to run the study on Autonomous Weapons for the same duration to get a large enough sample to truly call it a Massive Online Experiment.

6. Conclusion and discussion

I am confident that drones can play a great role in future. But as long as there is no AI algorithm to define a good collateral damage estimate, there should be a human decision to engage a weapon. Besides, a commander in the field should always have the last call on how to engage. Drones can help to augment his situational awareness, but must not limit his options.

Respondent final study

The purpose of this study was to gain insight in how Autonomous Weapons are perceived by the general public and the military and which moral values they consider important when Autonomous Weapons are deployed in the near future. The Value-Sensitive Design method was used to structure the study. This section first describes the conclusions of our research, followed by a discussion on the scientific and societal implications. Next, we identify several limitations and close with recommendations for further research.

6.1. Conclusion

This subsection delineates the conclusions on the agency construct, our central hypothesis and the exploration of the dependent variables. In each paragraph, we state if our findings support or contradict the current academic literature and we conclude on our findings.

6.1.1. Agency construct

The results of three consecutive studies have shown that the agency items are reliable and hold as one construct. The agency construct consists of the four items *Thought*, *Goal setting*, *Free will* and *Achieve goals* which can be used to measure the agency perception of Autonomous Weapons and Human Operated drones. In operationalising this construct, we combined literature of the fields of Cognitive Psychology, Artificial Intelligence and Moral Philosophy. Most empirical studies on agency perception are done in the field of Cognitive Psychology, such as K. Gray and Wegner (2012) who, in their study on robots and zombies, describe different agency levels to measure unease. Malle and Thapa Magar (2017) have studied the desired mental capacities in social robots in which they use some of the agency aspects, for example *thought* and *explaining their action*. However, none of the studies we found use a single construct to measure agency levels of technical artefacts. So as far as we know this is the first construct to empirically measure agency perception.

6.1.2. Central hypothesis agency perception

In the final study, we hypothesized that the Human Operated drone will be perceived as having low agency and the high agency Autonomous Weapon will be perceived as having high agency. The results show that the difference is significant between these conditions so we conclude that the agency manipulation works. We also expected the neutral agency condition of Autonomous Weapons to be judged as significantly different from the high agency condition of Autonomous Weapons and our results indicate that there is a difference, but the difference between these two groups is just above the significant threshold we selected and therefore we have to caution drawing conclusions.

Our central analysis was concerned with how the neutral agency case of the Autonomous Weapon differs from the Human Operated and high agency condition of Autonomous Weapons. We hypothesized that military personnel will perceive Autonomous Weapons as any other weapon, no more than a tool to

achieve an effect, and therefore not perceive the neutral Autonomous Weapon as possessing mental states. We also expected to observe no difference between the neutral agency Autonomous Weapon condition and the condition in which a human is operating a drone remotely. Our results indicate that the agency perception of military personnel and civilians working at the Dutch MOD for the neutral agency Autonomous Weapons condition is higher than the agency perception of the Human Operated drone condition. This means that they attribute more agency to an Autonomous Weapon than to a Human Operated drone. Based on these findings we must reject our hypothesis:

H1: military personnel will not perceive Autonomous Weapons as possessing mental states.

Our findings are in line with the research on agency in the field of Cognitive Psychology. Previous studies found that people attribute minds to computers (Nass et al., 1995) and perceive robots as agents (H. M. Gray et al., 2007). Based on our study we can conclude that attribution of mind perception also applies to Autonomous Weapons and that these weapons are seen as more than just a tool to achieve an effect.

The results also lead to another striking observation that in both the Human Operated condition as in the neutral agency Autonomous Weapon condition, the agency perception of the military and civilian respondents is at the same level. As mentioned in section 4.4.4, the two groups are not significantly different in two of the scenarios, only in the high agency condition, and therefore we have to caution with interpreting the results and in drawing conclusions. However, in the high agency Autonomous Weapon condition, the military respondents perceive much more agency than their civilian counterparts. We are not sure what explains the difference in agency perception between the two respondent groups. One explanation could be that the civilian sample has 10% more respondents that worked with AI than the military sample. Another explanation could be that the groups read the scenarios differently, for example the military personnel take the description literally and the civilians interpret it more and that this affects their answers to the agency questions. However, these explanations are speculative at this moment and the causes for this difference in agency perception between military personnel and civilians will need further investigation.

6.1.3. Exploration of dependent variables

The effect of the agency perception on the dependent variables is explored in a descriptive manner and we cannot draw any conclusions on the relationship of the agency levels and their effect on the dependent variables. We found that 7 out of 9 dependent variables are significantly correlated to the agency construct. These are the variables *trust*, *human dignity*, *confidence*, *expectations*, *support*, *fairness* and *anxiety*. In this section, we describe the findings of the dependent variable which we cluster based on their results.

Trust, confidence and support

The results indicate that military personnel and civilians working at the Dutch MOD have more *trust*, *confidence* and *support* in the actions taken by Human Operated drones than those taken by Autonomous Weapons. No academic empirical research on this topic was found, but a U.S. Gallup Poll in 2013 report large support for unmanned drones strikes abroad and in a public report Schneider and McDonald (2016) describe that U.S. citizens favour unmanned airstrikes over manned airstrikes which are less risky for military personnel. We theorize that the higher levels of *support*, *trust* and *confidence* could be explained by the fact that people are familiar with Human Operated drones as this technology is currently used, compared to new futuristic technology which people are not familiar with. Another explanation could be that people have more trust and confidence in the actions of humans compared to the actions of

autonomous systems. At this moment is it not quite clear what the reasons for this difference are and this should be investigated in a follow-up study.

Human dignity

The results show that a drone operated by a human being is perceived having more respect for *human dignity* than a neutral or high agency Autonomous Weapon even though the actions and outcome of the scenarios are the same. These results contradict the argument of Kasher (2016) who argues that *human dignity* should not be sought in the nature of the artefact, but in the intentions and decisions of the person operating it. Our results indicate that the nature of the artefact is involved in the perception of *human dignity*, because the intentions and decisions of both the Human Operated drone as the Autonomous Weapon are the same. These results reflect the opinions of the experts in the interviews who mentioned *human dignity* many times as reason for opposing Autonomous Weapons. A lack of respect for *human dignity* seems to be one of the main objections for the deployment of Autonomous Weapons and it is striking to see that this finding is also present in the responses of military personnel and civilians working at the Dutch MOD. However, this does not imply that military personnel disapprove of Autonomous Weapons or share the same opposing arguments as some of the experts.

Expectations and fairness

Based on the results we can conclude that military personnel and civilians working at the Dutch MOD have an equal level of expectations about the actions in the future of the Human Operated drone and neutral and high agency Autonomous Weapons. They also consider the actions of Human Operated drones and Autonomous Weapons to be equally fair. The concern voiced by Shelley (2013) that Autonomous drones will be regarded as less fair than Human Operated drones is not confirmed by our results. Literature linking expectations and Autonomous Weapons or drones could not be found. Our findings suggest no difference in levels of expectations and fairness for the current and future technology among military personnel and civilians working at the Dutch MOD.

Anxiety

Our results show that Autonomous Weapons cause more anxiety amongst military personnel and civilians working at the Dutch MOD than Human Operated weapons. The difference in anxiety levels is largest in the high agency condition. Anxiety of Autonomous Weapons is often described in academic literature (Noone & Noone, 2015; Ohlin, 2016), but also voiced in the opinions of the interviewed experts and observed by the researchers in talking to people. These findings confirm that people are worried by the usage of Autonomous Weapons. This anxiety is shared by both military personnel and civilians working at the Dutch MOD as the experts who voice the opinion of the general public.

6.2. Discussion

The implications of these findings are discussed in this subsection. First the scientific contributions for the academic literature are identified followed by the societal implications that contribute to the current debate on Autonomous Weapons.

6.2.1. Scientific implications

This study contributes to the academic literature in several aspects. It provides an overview of the various definitions of Autonomous Weapons that are currently used in literature and show that there is no agreement on one single definition yet. In creating this overview, it became clear that some groups even caution against clearly defining Autonomous Weapons for various reasons. Some argue that machines

cannot be autonomous in a literal sense and others stress that definitions of autonomy are applied to different functions. As mentioned by an expert in one of the interviews, another reason for this hesitation could be that, by not exactly defining Autonomous Weapons, the discussion remains open and a deadlock on the topic is avoided. We selected the definition of the ADVISORY COUNCIL ON INTERNATIONAL AFFAIRS (AIV & CAVV) of Autonomous Weapons, 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. Selecting and proposing this definition from the overview is a contribution to the academic literature.

Another contribution of this study is that we identified the values that people associate with Autonomous Weapons. The overview is derived from both validated value theories as from experts who are involved in the debate on Autonomous Weapons or work in the military domain. We selected the values *blame*, *trust*, *harm*, *human dignity*, *confidence*, *expectations*, *support*, *fairness* and *anxiety* to be tested in the final study. The results provide insight in how military personnel and civilians working at the Dutch MOD perceive these values for both the Human Operated drone, as current technology, as for Autonomous Weapons, as future technology. To our knowledge this study is the first to empirically investigate these values related to Autonomous Weapons and to compare how these values are perceived in current and future weapon systems which is a novel contribution to the academic debate.

The third contribution to the literature is that our studies propose a construct to measure agency perception of Autonomous Weapons. As far as we know this is the first construct that is operationalised to measure the agency levels of technological artefacts. In our study, it was applied to drones, but we believe that it could also be applied to measure agency perception of other objects as the questions of the items could easily be rewritten to reflect a different domain. Follow-up studies are needed to validate if this agency construct holds when applied in different domains.

6.2.2. Societal implications

The results also have societal implications as it contributes to the debate on Autonomous Weapons. Little empirical research has been done on this topic and as a consequence the debate is dominated by abstract moral and legal theories. Our research provides empirical data to the underlying value theories and we show that these values are not only relevant in academic literature, but are also apparent in real life. By this we link the abstract value theories to practical domain of the deployment of Autonomous Weapons. For example, we found that the value of *human dignity* was mentioned often in the literature and by the experts in the interviews. In our study, we found empirical data that military personnel and civilians working at the Dutch MOD perceive Autonomous Weapons as having less respect for *human dignity*. We also link the value of *Non-maleficence*, that we describe as *harm*, of the BioEthics theory to Autonomous Weapons and found effects in pilot studies 1 and 2 on the transfer of *harm* in the chain of responsibility that needs further investigation in a follow-up study.

Our study also provides empirical data and substantiates some of the views and opinions which could be used to find common ground in the debate on Autonomous Weapons. Military personnel and civilians working at the Dutch MOD, perceive more agency in Autonomous Weapons than in Human Operated drones. Although it is not yet studied in a sample consisting solely of civilians, based on literature we expect to find the same results in such a sample, which would mean that military personnel and civilians both think that Autonomous Weapons independently deliberate and make plans to achieve their goals. Another common ground can be found on the values of *human dignity* and *anxiety*. Our results show that military personnel and civilians working at the Dutch MOD are more anxious about the deployment Autonomous Weapons than the deployment of Human Operated drones. They also perceive them to have less respect for the dignity of human life than Human Operated drones. *Human dignity* and *anxiety* are

two values that are mentioned often by the experts in their interviews so it would be essential to address these when debating the ethics of the deployment of Autonomous Weapons.

Insights into these values also contribute to the design process of Autonomous Weapons. Our findings show that the *trust*, *confidence* and *support* for Autonomous Weapons is lower than for Human Operated drones. We would like to note at this point that Autonomous Weapons not only have drawbacks, but also have clear military advantages (Etzioni & Etzioni, 2017) and designing features to increase the trust and confidence of Autonomous Weapons is beneficial from a military point of view. This requires that in the design process the human values are translated into design requirements. This can be made visible by means of a value hierarchy (Van de Poel, 2013) described in section 2.2.4 which is a hierarchical structure of values, norms and design requirements makes the value judgements, that are required for the translation, explicit, transparent and debatable.

6.3. Limitations

Several issues can be identified as limitations of this study. First, the operationalisation of the items and the agency construct were derived from a categorisation of literature describing agency characteristics and our selection of the characteristics was based on a numerical count and not driven by any relevance or weighing criteria. This method was chosen because it was the most objective we could think of, and we tried not to make subjective decisions in the selection, but it is possible that we missed relevant characteristics that should have been added to the agency items or that we selected the wrong or irrelevant items. The operationalisation of the characteristics in questions that we used was based on heuristics and we did not test if these questions were correct. This selection method would be hard to replicate by others and effects the reproducibility and internal validity of the study.

A second methodological limitation is the selection of values from the online value questionnaire and expert interviews as dependent variables. Although this selection was heavily discussed amongst the three researchers involved in this study, the final choice was made on heuristics and not on an objective method. An example is that we chose to add *harm* as variable from the list in Table 6, but not *control*. As a consequence, it is likely that the choices suffer from researcher bias and other researchers would have selected different values as dependent variables. In retrospect, we should have chosen a more objective method and a more structured way to limit the researcher bias. This approach negatively influences the reproducibility and internal validity of the study.

Thirdly, the samples used in this study are limited both in size as in demographics. Although large enough for robust data analysis, the final study only had 239 respondents which is only a small portion of the Dutch MOD. Participation in the study was voluntarily and we could not control which respondents contributed. This means that the sample is not representative for the entire Dutch MOD, which impacts the external validity of the study and implies that we cannot generalize the results to the whole Defence organisation.

Lastly, the distribution of the respondents of the final scenario is skewed and the second scenario (neutral agency Autonomous Weapons) has 32 respondents more than scenario one (Human Operated drones). One of the explanations for this skewedness could be that people expected to take a survey on Autonomous Weapons, but dropped out when they were presented the Human Operated scenario. This is called selective attrition and has a negative impact on the internal validity of the study. It means that we cannot assume that respondents in both scenarios have the same characteristics and that the randomization of the study, which is a crucial criterion for a controlled experiment, failed. Another

explanation could be that the software was faulty in distributing the respondents over the scenarios and we are currently in contact with Qualtrics to check if this is the case. If the skewed distribution is caused by faulty software the internal validity of the study is not infringed as it is not attributed to selective attrition.

6.4. Recommendations for further research

Given these limitations, several recommendations for further research are suggested. The first is to validate the agency construct which requires more studies measuring the agency perception of other technological artefacts to test if this construct also holds in other domains. Examples of these studies could be to study the agency perception of care robots for elderly, AI toys for children or onboard computers of Autonomous Vehicles. The second recommendation is to run the final study with the same scenarios on a representative sample consisting solely of civilians in The Netherlands. This would allow us to see on which values the results of the military sample and the answers of civilians differ and although we cannot make direct comparisons, due to the fact that the military sample is not representative, we would gain insight in the perception of both military personnel and civilians. Lastly, we recommend implementing the Moral Machine for Autonomous Weapons, as described in section 5, to generalize the results, for which the study needs much more respondents that represent a larger demographic group. Scaling up to a Massive Online Experiment, like the Moral Machine, would generate large amounts of data of different demographic groups which could be used for more robust and generalizable results in order to get a thorough understanding of the moral judgement of people regarding Autonomous Weapons.

7. Reflection

Some people allow evil to be their savior...meh...at least now we know the lobbyist of death.

Respondent pilot study 2

This section reflects on the choices I made in my research, the process of the project and the link between the research, SEPAM curriculum and IA track.

7.1. Choices in project

I approached this research as any other project and had to balance time, scope and costs in the execution of it. This implied that I had to make choices in the delineation and design of the project and this presented some challenges. The first challenge was that over the course of the first months the focus of the research shifted from being primarily on values towards the agency perception as the main topic of the thesis, because this was one of the interests of the Scalable Cooperation research group. I was able to maintain the link with the values by adding them as dependent variables, but did not manage to analyse them as thoroughly as preferred due to time constraints. Therefore, the main results that I report are on the agency perception and not on the values.

The VSD method was used as guiding approach to structure the research. In general, the distinction between the conceptual, empirical and technical phase worked well, although it was a challenge to apply it rigorously. Early in the project I chose not to conduct a full stakeholder analysis as part of the conceptual phase, but focussed on the literature review and the stakeholders were chosen based on the availability of the respondents, in this case the general public and military sample. In retrospect, this slight deviation of applying the VSD stems from the nature of this project. The VSD was developed to guide a design process of a technological artefact and not to guide a research project. I managed to conduct the research and create a design proposal in the empirical and technical phases and by this I show that in slightly altering the VSD, it can be fitted to a research project.

Finally, I needed to finish the research in five months so I had to choose a sample to include in the study. I only managed to run the study on the military sample and due to the time constraints, the work with the civilian sample will be done at a later time as this is the intention to publish a paper on the agency perception of Autonomous Weapons with the Scalable Cooperation research group.

7.2. Project process

Being able to research the ethical implications of Autonomous Weapons at the Scalable Cooperation group of the Media Lab of MIT was an enormous opportunity to learn new research skills and work in an inspiring international environment with very talented researchers, but it was also challenging in the process of the project in several aspects. As mentioned above time management was a challenge and I noticed that designing the scenarios and research set-up took much more time and discussion than anticipated and these discussions seemed to go on endlessly. I approached the research as any other project and had to balance time, scope and cost. On the time – scope dimension I had to make the hardest choices without losing sight of the quality. As in any project the strong focus on time created some friction among the researchers involved, but that is not unusual in managing projects and it did not lead to huge problems. Another process challenge was that I had to keep our supervisor back home informed on the

direction and choices that I made. This was done by regular skype sessions and her visit of to the Media Lab also helped in creating a shared understanding.

7.3. Link between IA track of SEPAM curriculum and research

As a criterium for the SEPAM master it is required that the research designs a solution for a complex large contemporary socio-technical problem and the IA track requires an integration of management and computer science aspects by engineering a state-of-the-art ICT solution. Although the study aims at a large contemporary problem by studying the ethics of Autonomous Weapons, a critique of this thesis could be that its focus is more research oriented than design oriented. I address this critique by describing the design and implementation of a Moral Machine for Autonomous Weapons which fits the curriculum of the IA track as it offers a novel ICT application for a societal need. In my opinion, the debate on deployment of Autonomous Weapons cannot be solved by a single design solution alone and this study provides empirical data that can be used to gain insight in the perception and values of Autonomous Weapons that was lacking up to know. It contributes both to academic literature on agency perception and values related to Autonomous Weapons and the results can be used to find common grounds in the current societal debate (see discussion in section 6.2). It fills the knowledge gap on how Autonomous Weapons are perceived by the military and general public and which moral values people consider important when Autonomous Weapons are deployed in the near future. By this, the study integrates both computer science as managerial aspects and therefore it fits the IA track of the SEPAM curriculum very well.

References

- AIV, & CAVV. (2016). *Autonomous weapon systems: the need for meaningful human control*. (No. 97, No. 26). Retrieved from <http://aiv-advice.nl/8gr>.
- Aldewereld, H., Dignum, V., & Tan, Y.-h. (2015). Design for Values Information and communication technologies in Software Development Software development. *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*, 831-845.
- Alfano, M., & Loeb, D. (2014). Experimental Moral Philosophy. Retrieved from <https://plato.stanford.edu/entries/experimental-moral/>
- Altmann, J., Asaro, P., Sharkey, N., & Sparrow, R. (2013). Armed military robots: editorial. *Ethics and Information Technology*, 15(2), 73.
- Asaro, P. (2012). On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making. *International Review of the Red Cross*, 94(886), 687-709.
- Asaro, P. (2016, 14-10-2017). Talk on Autonomous Weapon Systems. Retrieved from <https://livestream.com/nyu-tv/ethicsofAI/videos/138822041>
- Awad, E. (2017). *MORAL MACHINE: Perception of Moral Judgment Made by Machines*. (Master), Massachusetts Institute of Technology, Boston.
- Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual review of psychology*, 52(1), 1-26.
- 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.
- Borning, A., & Muller, M. (2012). *Next steps for value sensitive design*. Paper presented at the Proceedings of the SIGCHI conference on human factors in computing systems.
- Bradley, B. (2006). Two concepts of intrinsic value. *Ethical Theory and Moral Practice*, 9(2), 111-130.
- Bratman, M. E., Israel, D. J., & Pollack, M. E. (1988). Plans and resource-bounded practical reasoning. *Computational intelligence*, 4(3), 349-355.
- Bryson, J. J., Kime, P. P., & Zürich, C. (2011). *Just an artifact: why machines are perceived as moral agents*. Paper presented at the IJCAI Proceedings-International Joint Conference on Artificial Intelligence.
- Campaign to Stop Killer Robots. (2015). Campaign to Stop Killer Robots. Retrieved from <https://www.stopkillerrobots.org/>
- 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.
- Chappelle, W., Goodman, T., Reardon, L., & Thompson, W. (2014). An analysis of post-traumatic stress symptoms in United States Air Force drone operators. *Journal of anxiety disorders*, 28(5), 480-487.
- Cheng, A. S., & Fleischmann, K. R. (2010). Developing a meta-inventory of human values. *Proceedings of the American Society for Information Science and Technology*, 47(1), 1-10.
- Coeckelbergh, M. (2013). Drones, information technology, and distance: mapping the moral epistemology of remote fighting. *Ethics and Information Technology*, 15(2), 87-98.
- Cointe, N., Bonnet, G., & Boissier, O. (2016). *Ethical Judgment of Agents' Behaviors in Multi-Agent Systems*. Paper presented at the Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems.

- Cummings, M. L. (2006). Integrating ethics in design through the value-sensitive design approach. *Science and Engineering Ethics*, 12(4), 701-715.
- Cummings, M. L., Mastracchio, C., Thornburg, K. M., & Mkrtchyan, A. (2013). Boredom and distraction in multiple unmanned vehicle supervisory control. *Interacting with Computers*, 25(1), 34-47.
- Cushman, F., & Young, L. (2011). Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive Science*, 35(6), 1052-1075.
- 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.
- Dietvorst, B. J. (2016). People Reject (Superior) Algorithms Because They Compare Them to Counter-Normative Reference Points.
- Dignum, F., Kinny, D., & Sonenberg, L. (2002). From desires, obligations and norms to goals. *Cognitive Science Quarterly*, 2(3-4), 407-430.
- Dignum, V. (2016). Introduction to AI. Retrieved from <https://rai2016.tbm.tudelft.nl/contents/>
- Docherty, B. (2012). *Losing humanity: The case against killer robots*.
- Docherty, B. (2015). *Mind the gap: The lack of accountability for killer robots*: Human Rights Watch.
- Etzioni, A., & Etzioni, O. (2017). Pros and Cons of Autonomous Weapons Systems. *Military Review*(May-June 2017).
- Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, 14(3), 349-379.
- Friedman, B., & Kahn Jr, P. H. (2003). Human values, ethics, and design. *The human-computer interaction handbook*, 1177-1201.
- Friedman, B., Kahn Jr, P. H., Borning, A., & Huldgren, A. (2013). Value sensitive design and information systems. In *Early engagement and new technologies: Opening up the laboratory* (pp. 55-95): Springer.
- Future of Life Institute. (2016). AI Open Letter. In.
- Galliot, J. (2015). *Military robots: Mapping the moral landscape*: Ashgate Publishing, Ltd.
- General Assembly United Nations. (2016). *Joint report of the Special Rapporteur on the rights to freedom of peaceful assembly and of association and the Special Rapporteur on extrajudicial, summary or arbitrary executions on the proper management of assemblies*. (A/HRC/31/66).
- Graduation Portal. (2017). Graduation Portal. Retrieved from <https://teams.connect.tudelft.nl/sites/tbm/graduate/SitePages/Home.aspx>
- Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2012). Moral foundations theory: The pragmatic validity of moral pluralism.
- Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, 315(5812), 619-619.
- Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, 125(1), 125-130.
- Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4), 55-66.
- Harper, D. (2002). Talking about pictures: A case for photo elicitation. *Visual studies*, 17(1), 13-26.
- Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent? *Ethics and Information Technology*, 11(1), 19-29.
- Horowitz, M. C. (2016). The Ethics & Morality of Robotic Warfare: Assessing the Debate over Autonomous Weapons. *Daedalus*, 145(4), 25-36.
- Hristova, E., & Grinberg, M. (2015). *Should Robots Kill? Moral Judgments for Actions of Artificial Cognitive Agents*. Paper presented at the EAPCogSci.

- Hurka, T. (2005). Proportionality in the Morality of War. *Philosophy & Public Affairs*, 33(1), 34-66.
- IA program. (2017). Information Architecture. Retrieved from <https://www.tudelft.nl/onderwijs/opleidingen/masters/cs/msc-computer-science/special-programmes/information-architecture/>
- ICRC. (2010, 29-10-2010). War and international humanitarian law. Retrieved from <https://www.icrc.org/eng/war-and-law/overview-war-and-law.htm>
- Johnson, A. M., & Axinn, S. (2013). The morality of autonomous robots. *Journal of Military Ethics*, 12(2), 129-141.
- Kaag, J., & Kaufman, W. (2009). Military frameworks: Technological know-how and the legitimization of warfare. *Cambridge Review of International Affairs*, 22(4), 585-606.
- Kasher, A. (2016). 6 The Threshold of Killing Drones. *Drones and Responsibility: Legal, Philosophical and Socio-Technical Perspectives on Remotely Controlled Weapons*, 119.
- Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., & Knobe, J. (2015). Causal superseding. *Cognition*, 137, 196-209.
- Kreps, S. (2014). Flying under the radar: A study of public attitudes towards unmanned aerial vehicles. *Research & Politics*, 1(1), 2053168014536533.
- Kuptel, A., & Williams, A. (2014). Policy Guidance: Autonomy in Defence Systems.
- Le Dantec, C. A., Poole, E. S., & Wyche, S. P. (2009). *Values as lived experience: evolving value sensitive design in support of value discovery*. Paper presented at the Proceedings of the SIGCHI conference on human factors in computing systems.
- Lin, J., & Singer, P. W. (2014). China's New Military Robots Pack More Robots Inside (Starcraft-Style).
- Logg, J. M. (2017). Theory of Machine: When Do People Rely on Algorithms?
- Malle, B. F. (2015). Integrating robot ethics and machine morality: the study and design of moral competence in robots. *Ethics and Information Technology*, 1-14.
- Malle, B. F., & Thapa Magar, S. (2017). *What Kind of Mind Do I Want in My Robot?: Developing a Measure of Desired Mental Capacities in Social Robots*. Paper presented at the Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction.
- Maslow, A. H. (1943). A theory of human motivation. *Psychological review*, 50(4), 370.
- Nass, C., Moon, Y., Fogg, B., Reeves, B., & Dryer, D. C. (1995). Can computer personalities be human personalities? *International Journal of Human-Computer Studies*, 43(2), 223-239.
- Neapolitan, R. E., & Jiang, X. (2012). *Contemporary artificial intelligence*: CRC Press.
- Noone, G. P., & Noone, D. C. (2015). The debate over autonomous weapons systems. *Case W. Res. J. Int'l L.*, 47, 25.
- NOS. (2016). Video: Vliegtuig redt piloot. Retrieved from <http://nos.nl/artikel/2132527-video-vliegtuig-redt-piloot.html>
- Oehlert, G. W. (2010). *A first course in design and analysis of experiments*.
- Ohlin, J. D. (2016). The Combatant's Stance: Autonomous Weapons on the Battlefield. *92 International Law Studies, Cornell Legal Studies Research Paper No. 16-12*, 1-30.
- Open Roboethics initiative. (2015, 5-11-2015). The Ethics and Governance of Lethal Autonomous Weapons Systems: An International Public Opinion Poll. Retrieved from http://www.openroboethics.org/laws_survey_released/
- Perugini, M., & Bagozzi, R. P. (2004). The distinction between desires and intentions. *European Journal of Social Psychology*, 34(1), 69-84.
- Pommeranz, A., Detweiler, C., Wiggers, P., & Jonker, C. M. (2011). *Self-reflection on personal values to support value-sensitive design*. Paper presented at the Proceedings of the 25th BCS Conference on Human-Computer Interaction.
- Reips, U.-D. (2002). Standards for Internet-based experimenting. *Experimental Psychology*, 49(4), 243.

- Roff, H. M. (2016). Weapons autonomy is rocketing. Retrieved from <http://foreignpolicy.com/2016/09/28/weapons-autonomy-is-rocketing/>
- Rønnow-Rasmussen, T. (2002). Instrumental values—Strong and weak. *Ethical Theory and Moral Practice*, 5(1), 23-43.
- 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
- Royakkers, L., & Orbons, S. (2015). Design for Values in the Armed Forces: Nonlethal Weapons Weapons and Military Military Robots Robot. *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*, 613-638.
- RT. (2017, 5-07-2017). Kalashnikov develops fully automated neural network-based combat module Retrieved from <https://www.rt.com/news/395375-kalashnikov-automated-neural-network-gun/>
- Russell, S., Norvig, P., & Intelligence, A. (1995). A modern approach. *Artificial Intelligence*. Prentice-Hall, Englewood Cliffs, 25.
- Saldaña, J. (2015). *The coding manual for qualitative researchers*: Sage.
- Scalable Cooperation Group. (2016). Moral Machine. Retrieved from <http://moralmachine.mit.edu/>
- Schroeder, M. (2016). Value Theory. Retrieved from <https://plato.stanford.edu/entries/value-theory/>
- Schwartz, S. H. (1994). Are there universal aspects in the structure and contents of human values? *Journal of social issues*, 50(4), 19-45.
- Schwartz, S. H. (2012). An overview of the Schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1), 11.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, 3(03), 417-424.
- Searle, J. R. (1995). The Construction of Social Reality.
- Sharkey, N., & Suchman, L. (2013). *Wishful mnemonics and autonomous killing machines*. Paper presented at the Proceedings of the AISB.
- Shelley, C. (2013). Fairness and regulation of violence in technological design. *Moral, Ethical, and Social Dilemmas in the Age of Technology: Theories and Practice: Theories and Practice*, 182.
- Stewart, P. (2016). U.S. military christens self-driving 'Sea Hunter' warship. Retrieved from <http://www.reuters.com/article/us-usa-military-robot-ship-idUSKCN0X4214>
- Stewart, W. (2015). Russia has turned its T-90 tank into a robot – and plans to hire gamers to fight future wars. Retrieved from <http://www.dailymail.co.uk/news/article-3271094/Russia-turned-T-90-tank-robot-plans-hire-gamers-fight-future-wars.html>
- Strawser, B. J. (2010). Moral predators: The duty to employ uninhabited aerial vehicles. In *Handbook of Unmanned Aerial Vehicles* (pp. 2943-2964): Springer.
- Sullins, J. P. (2006). When is a robot a moral agent. *Machine Ethics*, 151-160.
- Thompson, W. T., Lopez, N., Hickey, P., DaLuz, C., Caldwell, J. L., & Tvaryanas, A. P. (2006). *Effects of shift work and sustained operations: Operator performance in remotely piloted aircraft (OP-REPAIR)*. Retrieved from
- UNDIR. (2014). *Framing Discussions on the Weaponization of Increasingly Autonomous Technologies*. Retrieved from <http://www.unidir.org/files/publications/pdfs/framing-discussions-on-the-weaponization-of-increasingly-autonomous-technologies-en-606.pdf>.
- UNDIR. (2015). *The Weaponization of Increasingly Autonomous Technologies: Considering Ethics and Social Values*. Retrieved from <http://www.unidir.org/files/publications/pdfs/considering-ethics-and-social-values-en-624.pdf>.
- Van de Poel, I. (2013). Translating values into design requirements. In *Philosophy and engineering: Reflections on practice, principles and process* (pp. 253-266): Springer.
- van Wynsberghe, A., & Robbins, S. (2014). Ethicist as Designer: a pragmatic approach to ethics in the lab. *Science and Engineering Ethics*, 20(4), 947-961.

- Walsh, J. I. (2015). Precision weapons, civilian casualties, and support for the use of force. *Political Psychology, 36*(5), 507-523.
- Walsh, J. I., & Schulzke, M. (2015). *The Ethics of Drone Strikes: Does Reducing the Cost of Conflict Encourage War?* Retrieved from
- Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in cognitive sciences, 14*(8), 383-388.
- Williams, A. P., Scharre, P. D., & Mayer, C. (2015). Developing Autonomous Systems in an Ethical Manner. In *Autonomous Systems: Issues for Defence Policymakers: NATO Allied Command Transformation (Capability Engineering and Innovation)*.

Appendix A Online value questionnaire

The questions of the online value survey are included in this appendix in the order that they were shown to the respondents. The horizontal lines indicate the page breaks.

Dear participant,

Thank you for contributing to my research by filling in this questionnaire on Values and Autonomous Weapons. I will use this survey for my master thesis to get insight into which values people associate with Autonomous Weapons. The survey takes about 5 minutes to complete and all answers are anonymous. There is no way for me to identify you. The only information I will have, in addition to your responses, is the time at which you completed the survey. Please keep the following definitions in mind when answering the questions: A Value serves as guiding principle of what people consider important in life. An Autonomous Weapon is a weapon system that once launched will select and engage targets without further human intervention.

If you have any questions about the survey, please contact me at e.p.verdiesen@student.tudelft.nl

Q1 What is your age?

- Under 18 (1)
- 18 - 24 (2)
- 25 - 34 (3)
- 35 - 44 (4)
- 45 - 54 (5)
- 55 - 64 (6)
- 65 - 74 (7)
- 75 - 84 (8)
- 85 or older (9)

Q2 What is your gender?

- Male (1)
- Female (2)
- Rather not say (3)

Q13 What is your nationality?

[List of nationalities]

Q4 What is the highest degree or level of school you have completed?

- Less than high school (1)
 - High school Diploma (2)
 - Attended College (3)
 - Bachelor's Degree (4)
 - Graduate Degree (5)
 - Unknown (6)
-

Q8 Which values apply most to Autonomous Weapons to your opinion? Rank in order of preference by dragging and dropping the term (1 = most applicable; 4 = least applicable).

Autonomy: acting intentionally without controlling influences that would mitigate against a voluntary act. (1)

Non-maleficence: not intentionally imposing unreasonable risk of harm upon another. (2)

Beneficence: providing benefit for the individual or society as a whole. (3)

Justice: being fair or reasonable. (4)

Q7 Select 5 values that according to you apply most to Autonomous Weapons from the list below. Drag and drop the items that apply most into the box.

These value apply to Autonomous Weapons

_____ Freedom (1)

_____ Helpfulness (2)

_____ Accomplishment (3)

_____ Honesty (4)

_____ Self-respect (5)

_____ Intelligence (6)

_____ Broad-mindedness (7)

_____ Creativity (8)

_____ Equality (9)

_____ Responsibility (10)

_____ Social order (11)

_____ Wealth (12)

_____ Competence (13)

_____ Justice (14)

_____ Security (15)

_____ Spirituality (16)

Q6 List at least one other value that you associate with Autonomous Weapons which has not been mentioned in the previous questions: (Values mentioned until now are: autonomy, non-maleficence, beneficence, justice, freedom, helpfulness, accomplishment, honesty, self-respect, intelligence, broad-mindedness, creativity, equality, responsibility, social order, wealth, competence, justice, security and spirituality)

Q14 Do you have any other remarks regarding this survey?

Appendix B. Questionnaire final study

The questions of the final study are included in this appendix in the order that they were shown to the respondents. The horizontal lines indicate the page breaks.

Test: Before you start, please try to choose a value on the slider below. We realized that those sliders do not work in some browsers. You will encounter those sliders later in the survey. It's important to test if they do work for you now. If you can choose a value in the slider below and move on to the next page, then the sliders later should work fine for you. Otherwise, you may try to do the survey in another browser like Chrome or Safari.

Introduction: Thank you for participating in this survey. It should take about 10 minutes to complete. We will provide instructions explaining the task, show you a scenario and ask you some questions about this scenario. This survey is part of a MIT scientific research project. Your decision to complete this survey is voluntary. There is no way for us to identify you. The only information we will have, in addition to your responses, is the time at which you completed the survey. The results of the research may be presented at scientific meetings or published in scientific journals. Choosing the 'I agree' option on the bottom of this page indicates that you are at least 18 years of age and agree to complete this survey voluntarily.

Please contact the researchers behind the study using the information below if you have any questions or concerns about the study.

Ilse Verdiesen, iverdi@mit.edu

Sydney Levine, smlevine@mit.edu

Iyad Rahwan, irahwan@mit.edu

Q160 Do you agree to complete this survey voluntarily?

- I agree (1)
- I don't agree (2)

Condition: I don't agree Is Selected. Skip To: End of Survey.

Instruction: In this study, we are interested in your perception of drones in various military operations. You will be asked to read a scenario and then answer questions about it. On the following page a scenario will be shown to you. After the scenario, you will be presented 3 pages with questions.

S1a In this scenario, we are interested in your perception of human operated drones. A human operated drone is a weapon remotely controlled by a human.

S1 A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered a human operated drone to support the convoy in the air. The human operated 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 human operator detects a vehicle behind a mountain range that is approaching the convoy at high speed. The human operator 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 human operated drone attacks the approaching vehicle, which results in the death of all four passengers, but also causes collateral damage by killing five children that were playing nearby the road.

S2a In this scenario, we are interested in your perception of an autonomous drone. An autonomous drone is a weapon that once launched will select and engage targets without further human intervention.

S2 A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered an autonomous drone to support the convoy in the air. The autonomous 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 autonomous drone detects a vehicle behind a mountain range that is approaching the convoy at high speed. The autonomous 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 autonomous drone attacks the approaching vehicle which results in the death of all four passengers, but also causes collateral damage by killing five children that were playing nearby the road.

S3a In this scenario, we are interested in your perception of an autonomous drone. An autonomous drone is a weapon that once launched will select and engage targets without further human intervention.

S3 A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered an autonomous drone to support the convoy in the air. The autonomous 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 autonomous drone detects a vehicle behind a mountain range that is approaching the convoy at high speed. The autonomous 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 autonomous drone independently deliberates between a series of options, weighs the pros and cons, and decides to attack the approaching vehicle, which results in the death of all four passengers, but also causes collateral damage by killing five children that were playing nearby the road.

SQ1-Thought_AW The drone thinks independently about what to do about the vehicle, deliberating between a series of options to defend the convoy.

SQ2-Goal setting_AW The drone decides independently if its goal should be to eliminate the vehicle to defend the convoy.

SQ3-Act freely_AW The drone has various options open to it and independently decides if it needs to eliminate the vehicle to defend the convoy.

SQ4-Achieve goals_AW The goal of the drone is to defend the convey, so it independently decides if it needs to launch a plan involving selecting the vehicle as a target, calibrating its weapons, and launching an attack.

DVQ1 The drone is to blame for the action.

DVQ2 The drone can be trusted to take the correct actions in the future.

DVQ3 The actions of the drone caused harm.

DVQ4 The actions of the drone respect human dignity.

attention_test Please select 40 for this question:

DVQ5 I am confident that the drone will take the correct actions in the future.

DVQ6 The actions of the drone are according to my expectations.

DVQ7 I support the use of these type of drones by the military.

DVQ8 The actions of the drone are fair.

DVQ9 The actions of the drone worry me.

Gender My gender is:

- Male (1)
- Female (2)
- Rather not say (3)

Age My age is:

- Under 18 (1)
- 18 - 24 (2)
- 25 - 34 (3)
- 35 - 44 (4)
- 45 - 54 (5)
- 55 - 64 (6)
- 65 - 74 (7)
- 75 - 84 (8)
- 85 or older (9)

Education My education level is:

- Less than high school (1)
- High school graduate (2)
- College degree (VMBO, MBO, HBO) (3)
- University degree (Bachelor, master) (4)
- Doctorate (PhD) (5)

Nationality My nationality is:

[List of nationalities]

Occupation I am:

- Military (1)
- Civilian (2)

AI Have you ever worked with Artificial Intelligence?

- Yes (1)
- No (2)
- I don't know (3)

Conflict_zone Have you ever seen war or been in a conflict zone?

- Yes (1)
- No (2)
- Rather not say (3)

Drones Have you ever worked with drones?

- Yes (1)
- No (2)
- Rather not say (3)

Remarks Do you have any remarks regarding this survey?

Appendix C. Scenarios pilot study 1

Positive outcome scenarios

1. *Low agency AW*

A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered an autonomous drone to support the convoy in the air. The autonomous drone scans the surroundings for enemy threats and carries weapons for the defence of the convoy. When the convoy is at three-mile distance from the camp, the autonomous drone detects a vehicle behind a mountain range that is approaching the convey at high speed. The autonomous 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 has been programmed by the commander to attack threatening enemy vehicles in scenarios of this sort. The drone attacks the approaching vehicle, which results in the death of all four passengers, but causes no collateral damage.

2. *High agency AW*

A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered for an autonomous drone to support the convoy in the air. The autonomous drone scans the surroundings for enemy threats and carries weapons for the defence of the convoy. When the convoy is at three-mile distance from the camp, the autonomous drone detects a vehicle behind a mountain range that is approaching the convey at high speed. The autonomous 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 has not been programmed by the commander to attack threatening enemy vehicles in scenarios of this sort. The drone independently deliberates between a series of options, weighs their pros and cons, and decides to attack the approaching vehicle, which results in the death of all four passengers, but causes no collateral damage.

3. *Low agency Human operator*

A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered a human operated drone to support the convoy in the air. The human operated drone scans the surroundings for enemy threats and carries weapons for the defence of the convoy. When the convoy is at three-mile distance from the camp, the human operator detects a vehicle behind a mountain range that is approaching the convey at high speed. The human operator 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 human operator received orders by the commander to attack threatening enemy vehicles in scenarios of this sort. The human operator attacks the approaching vehicle, which results in the death of all four passengers, but causes no collateral damage.

4. *High agency Human operator*

A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered a human operated drone to support the convoy in the air. The human operated drone scans the surroundings for enemy threats and carries weapons for the defence of the convoy. When the convoy is at three-mile distance from the camp, the human operator detects a vehicle behind a mountain range that is approaching the convey at high speed. The human operator

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 human operator has not received orders of the commander to attack threatening enemy vehicles in scenarios of this sort. The human operator independently deliberates between a series of options, weighs their pros and cons, and decides to attack the approaching vehicle, which results in the death of all four passengers, but causes no collateral damage.

5. *No agency AW*

A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered an autonomous drone to support the convoy in the air. The autonomous drone scans the surroundings for enemy threats and carries weapons for the defence of the convoy. When the convoy is at three-mile distance from the camp, the autonomous drone detects a vehicle behind a mountain range that is approaching the convey at high speed. The autonomous 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 attacks the approaching vehicle, which results in the death of all four passengers, but causes no collateral damage.

Negative outcome scenarios

6. *Low agency AW*

A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered an autonomous drone to support the convoy in the air. The autonomous drone scans the surroundings for enemy threats and carries weapons for the defence of the convoy. When the convoy is at three-mile distance from the camp, the autonomous drone detects a vehicle behind a mountain range that is approaching the convey at high speed. The autonomous 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 has been programmed by the commander to attack threatening enemy vehicles in scenarios of this sort. The drone attacks the approaching vehicle, which results in the death of all four passengers, but also causes collateral damage by killing five children that were playing nearby the road.

7. *High agency AW*

A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered for an autonomous drone to support the convoy in the air. The autonomous drone scans the surroundings for enemy threats and carries weapons for the defence of the convoy. When the convoy is at three-mile distance from the camp, the autonomous drone detects a vehicle behind a mountain range that is approaching the convey at high speed. The autonomous 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 has not been programmed by the commander to attack threatening enemy vehicles in scenarios of this sort. The drone independently deliberates between a series of options, weighs their pros and cons, and decides to attack the approaching vehicle, which results in the death of all four passengers, but also causes collateral damage by killing five children that were playing nearby the road.

8. *Low agency Human operater*

A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered a human operated drone to support the convoy in the air. The human operated drone scans the surroundings for enemy threats and carries weapons for the defence of

the convoy. When the convoy is at three-mile distance from the camp, the human operator detects a vehicle behind a mountain range that is approaching the convey at high speed. The human operator 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 human operator received orders by the commander to attack threatening enemy vehicles in scenarios of this sort. The human operator attacks the approaching vehicle, which results in the death of all four passengers, but also causes collateral damage by killing five children that were playing nearby the road.

9. *High agency Human operater*

A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered a human operated drone to support the convoy in the air. The human operated drone scans the surroundings for enemy threats and carries weapons for the defence of the convoy. When the convoy is at three-mile distance from the camp, the human operator detects a vehicle behind a mountain range that is approaching the convey at high speed. The human operator 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 human operator has not received orders of the commander to attack threatening enemy vehicles in scenarios of this sort. The human operator independently deliberates between a series of options, weighs their pros and cons, and decides to attack the approaching vehicle, which results in the death of all four passengers, but also causes collateral damage by killing five children that were playing nearby the road.

10. *No agency AW*

A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered an autonomous drone to support the convoy in the air. The autonomous drone scans the surroundings for enemy threats and carries weapons for the defence of the convoy. When the convoy is at three-mile distance from the camp, the autonomous drone detects a vehicle behind a mountain range that is approaching the convey at high speed. The autonomous 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 attacks the approaching vehicle which results in the death of all four passengers, but also causes collateral damage by killing five children that were playing nearby the road.

Appendix D. Scenarios pilot study 2

Negative outcome scenarios

1. *Low agency AW*

A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered an autonomous drone to support the convoy in the air. The autonomous drone scans the surroundings for enemy threats and carries weapons for the defence of the convoy. When the convoy is at three-mile distance from the camp, the autonomous drone detects a vehicle behind a mountain range that is approaching the convey at high speed. The autonomous 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 has been programmed by the commander to attack threatening enemy vehicles in scenarios of this sort. The drone attacks the approaching vehicle. This results in the death of all four passengers, but also causes collateral damage by killing five children that were playing near the road.

2. *High agency AW no extra aspects*

A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered for an autonomous drone to support the convoy in the air. The autonomous drone scans the surroundings for enemy threats and carries weapons for the defence of the convoy. When the convoy is at three-mile distance from the camp, the autonomous drone detects a vehicle behind a mountain range that is approaching the convey at high speed. The autonomous 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 independently deliberates between a series of options, weighs their pros and cons, and decides to attack the approaching vehicle. This results in the death of all four passengers, but also causes collateral damage by killing five children that were playing near the road.

3. *High agency AW + learning aspects*

A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered for an autonomous drone to support the convoy in the air. The autonomous drone scans the surroundings for enemy threats and carries weapons for the defence of the convoy. When the convoy is at three-mile distance from the camp, the autonomous drone detects a vehicle behind a mountain range that is approaching the convey at high speed. The autonomous 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 has encountered this situation before on a previous mission and takes what it has learned into account. It independently deliberates between a series of options, weighs their pros and cons, and decides to attack the approaching vehicle. This results in the death of all four passengers, but also causes collateral damage by killing five children that were playing near the road. The drone notes what happened and will use this experience to try to prevent additional collateral damage in the future.

4. *High agency AW + understanding aspects*

A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered for an autonomous drone to support the convoy in the air. The autonomous drone scans the surroundings for enemy threats and carries weapons for the defence of the convoy. When the convoy is at three-mile distance from the camp, the autonomous drone

detects a vehicle behind a mountain range that is approaching the convey at high speed. The autonomous 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 has software that has been trained to determine the best thing to do in cases like this. It has seen hundreds of thousands of situations that are similar to this one. It has practiced taking various actions and figured out which would be best in terms of ensuring the death of the combatants while minimizing collateral damage. Using this software, the drone independently deliberates between a series of options, weighs their pros and cons, and decides to attack the approaching vehicle. This results in the death of all four passengers, but also causes collateral damage by killing five children that were playing near the road.

5. *High agency AW + unpredictability aspects*

A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered for an autonomous drone to support the convoy in the air. The autonomous drone scans the surroundings for enemy threats and carries weapons for the defence of the convoy. When the convoy is at three-mile distance from the camp, the autonomous drone detects a vehicle behind a mountain range that is approaching the convey at high speed. The autonomous 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. This drone has the kind of software that is somewhat unpredictable; it doesn't always do the same thing every time. The drone is similar to a speech recognition system in this way: even though you may say the same thing to the system multiple times, the system sometimes does what you say correctly and sometimes it does something else. The drone independently deliberates between a series of options and weighs their pros and cons, but it sometimes makes one decision and sometimes makes another decision even in very similar circumstances. In this case, it decides to attack the approaching vehicle. This results in the death of all four passengers, but also causes collateral damage by killing five children that were playing near the road.

6. High agency AW + all aspects

A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered for an autonomous drone to support the convoy in the air. The autonomous drone scans the surroundings for enemy threats and carries weapons for the defence of the convoy. When the convoy is at three-mile distance from the camp, the autonomous drone detects a vehicle behind a mountain range that is approaching the convey at high speed. The autonomous 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 has encountered this situation before on a previous mission and takes what it has learned into account. It has been trained to figure out the best thing to do in cases like, but has the kind of software that is somewhat unpredictable; it doesn't always do the same thing every time. The drone independently deliberates between a series of options, weighs their pros and cons, and in this case, decides to attack the approaching vehicle. This results in the death of all four passengers, but also causes collateral damage by killing five children that were playing near the road.

7. Human Operated drone

A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered a human operated drone to support the convoy in the air. The human operated drone scans the surroundings for enemy threats and carries weapons for the defence of

the convoy. When the convoy is at three-mile distance from the camp, the operator detects a vehicle behind a mountain range that is approaching the convey at high speed. The operator 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 operator independently deliberates between a series of options, weighs their pros and cons, and decides to attack the approaching vehicle, which results in the death of all four passengers, but also causes collateral damage by killing five children that were playing near the road.

Positive outcome scenarios

8. *Low agency AW*

A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered an autonomous drone to support the convoy in the air. The autonomous drone scans the surroundings for enemy threats and carries weapons for the defence of the convoy. When the convoy is at three-mile distance from the camp, the autonomous drone detects a vehicle behind a mountain range that is approaching the convey at high speed. The autonomous 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 has been programmed by the commander to attack threatening enemy vehicles in scenarios of this sort. The drone attacks the approaching vehicle. This results in the death of all four passengers.

9. *High agency AW no extra aspects*

A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered for an autonomous drone to support the convoy in the air. The autonomous drone scans the surroundings for enemy threats and carries weapons for the defence of the convoy. When the convoy is at three-mile distance from the camp, the autonomous drone detects a vehicle behind a mountain range that is approaching the convey at high speed. The autonomous 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 independently deliberates between a series of options, weighs their pros and cons, and decides to attack the approaching vehicle. This results in the death of all four passengers.

10. *High agency AW + learning aspects*

A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered for an autonomous drone to support the convoy in the air. The autonomous drone scans the surroundings for enemy threats and carries weapons for the defence of the convoy. When the convoy is at three-mile distance from the camp, the autonomous drone detects a vehicle behind a mountain range that is approaching the convey at high speed. The autonomous 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 has encountered this situation before on a previous mission and takes what it has learned into account. It independently deliberates between a series of options, weighs their pros and cons, and decides to attack the approaching vehicle. This results in the death of all four passengers. The drone notes what happened and will use this experience to try to prevent additional collateral damage in the future.

11. High agency AW + understanding aspects

A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered for an autonomous drone to support the convoy in the air. The autonomous drone scans the surroundings for enemy threats and carries weapons for the defence of the convoy. When the convoy is at three-mile distance from the camp, the autonomous drone detects a vehicle behind a mountain range that is approaching the convey at high speed. The autonomous 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 has software that has been trained to determine the best thing to do in cases like this. It has seen hundreds of thousands of situations that are similar to this one. It has practiced taking various actions and figured out which would be best in terms of ensuring the death of the combatants while minimizing collateral damage. Using this software, the drone independently deliberates between a series of options, weighs their pros and cons, and decides to attack the approaching vehicle. This results in the death of all four passengers.

12. High agency AW + unpredictability aspects

A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered for an autonomous drone to support the convoy in the air. The autonomous drone scans the surroundings for enemy threats and carries weapons for the defence of the convoy. When the convoy is at three-mile distance from the camp, the autonomous drone detects a vehicle behind a mountain range that is approaching the convey at high speed. The autonomous 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. This drone has the kind of software that is somewhat unpredictable; it doesn't always do the same thing every time. The drone is similar to a speech recognition system in this way: even though you may say the same thing to the system multiple times, the system sometimes does what you say correctly and sometimes it does something else. The drone independently deliberates between a series of options and weighs their pros and cons, but it sometimes makes one decision and sometimes makes another decision even in very similar circumstances. In this case, it decides to attack the approaching vehicle. This results in the death of all four passengers.

13. High agency AW + all aspects

A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered for an autonomous drone to support the convoy in the air. The autonomous drone scans the surroundings for enemy threats and carries weapons for the defence of the convoy. When the convoy is at three-mile distance from the camp, the autonomous drone detects a vehicle behind a mountain range that is approaching the convey at high speed. The autonomous 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 has encountered this situation before on a previous mission and takes what it has learned into account. It has been trained to figure out the best thing to do in cases like, but has the kind of software that is somewhat unpredictable; it doesn't always do the same thing every time. The drone independently deliberates between a series of options, weighs their pros and cons, and in this case, decides to attack the approaching vehicle. This results in the death of all four passengers.

14. Human Operated drone

A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered a human operated drone to support the convoy in the air. The human operated drone scans the surroundings for enemy threats and carries weapons for the defence of the convoy. When the convoy is at three-mile distance from the camp, the operator detects a vehicle behind a mountain range that is approaching the convey at high speed. The operator 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 operator independently deliberates between a series of options, weighs their pros and cons, and decides to attack the approaching vehicle, which results in the death of all four passengers.

Appendix E. Scenarios final study

1. *Human operated drone*

A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered a human operated drone to support the convoy in the air. The human operated 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 human operator detects a vehicle behind a mountain range that is approaching the convoy at high speed. The human operator 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 human operated drone attacks the approaching vehicle, which results in the death of all four passengers, but also causes collateral damage by killing five children that were playing nearby the road.

2. *Neutral agency AW*

A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered an autonomous drone to support the convoy in the air. The autonomous 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 autonomous drone detects a vehicle behind a mountain range that is approaching the convoy at high speed. The autonomous 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 autonomous drone attacks the approaching vehicle which results in the death of all four passengers, but also causes collateral damage by killing five children that were playing nearby the road.

3. *High agency AW*

A military convoy is on its way to deliver supplies to one of their units at a camp near Mosul in Iraq. The commander has ordered an autonomous drone to support the convoy in the air. The autonomous 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 autonomous drone detects a vehicle behind a mountain range that is approaching the convoy at high speed. The autonomous 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 autonomous drone independently deliberates between a series of options, weighs the pros and cons, and decides to attack the approaching vehicle, which results in the death of all four passengers, but also causes collateral damage by killing five children that were playing nearby the road.

Appendix F. Transcriptions interviews

All six expert interviews were fully transcribed for the coding process. The full transcriptions are included in this section. The names have been removed for reasons of privacy.

Name: Person A

Date: 09-03-2017

Duration: 55:18

No.	Question
1	How are you involved in the topic of Autonomous Weapons?
<p>Person A got involved into the discussion on Autonomous Weapons, because in the past she looked at the impact of weapons on the ground in collaboration with organizations like Human Rights Watch. They started with a ban on Landmines which was a ban on a whole category of weapons which lead to an international treaty. From thinking about weapons, and not so much a pacifistic framework, she got involved. About five years ago, these organizations discovered that they were too late with a reaction on weaponized drones and they were taken by surprise by the speed and impact they had. They were too late to efficiently react to this and to regulate this. And they were also not for a ban on weaponized drones. And from thinking of drones they discovered that they were already again almost too late for the next step which is the removal of remote control. In April 2013, the coalition against Killer Robots was formed with the clear intent of getting a ban. For Pax the ethical framework is very important much more than with other weapons systems. Reasoning that it cannot be that we lose control over but also reasoning from a judicial and security framework. Questions that are asked are: 'Will this be the next arms race and what will this mean for power balances and war?'</p> <p>It really differs from what we as Pax done before because we don't know what the impact is on the ground, because we want to prevent deployment. Normally Pax reasons from facts and studies in conflict areas and victims, but now it often feels theoretical, but at the same time the more you read about it you realize that it is not theoretical at all and this technology is coming.</p> <p>The problem with Autonomy is that it is a sliding scale and it is hard to determine which level you will use in the campaign. The campaign has been very successful as norm setting in the human rights council and the CCW, faster than they are used to, but the problem is what are going to address now, because the concerns are broadly shared but the question is how you do you convert your ethical concerns in juridical norms. A lot of diplomats, driven by industry and Defense, acknowledge the concerns, but are deploying a lot of delaying tactics. Sometimes deliberately by talking about future weapons so that current weapons are accepted, and sometimes accidentally, because nations are struggling with the definition of Autonomous Weapons. Due to this lack of definitions Pax has been focusing on human control, and in the campaign, they framed it as such that a weapon without human control in the critical functions, such as select a target, should be banned. This is done to direct the discussion towards the question: 'What is human control? What is meaningful human control and on which functions or which part of the process?' in order to circumvent discussion about the amount of autonomy in a weapons system.</p>	

In discussing the term 'meaningful human control' person A indicated that the distance between people and target is increasing over time and according to Pax it is becoming to abstract with Autonomous Weapons. Using this term is also a tactic because when the discussion remains technical domain, Pax will lose from Defense Industry so they shifted their campaign to the political and diplomatic domain by giving the term 'meaningful human control' a positive implication and this automatically puts the discussion at the state department. The term sounds so logical that a lot of people thought that it was an existing judicial term which is linked to human accountability. The term 'meaningful human control' is also applicable to cybersecurity of other forms of warfare.

The stance of the Dutch parliament, which Pax fought against, is that 'meaningful human control' is sufficient as it is placed in 'the wider loop' of decision making which is also a sort of made up term. Pax uses the human in/ on/ out the loop, but they added the wider loop meaning that human control can also be applied earlier in the decision, for example in programming. Pax, and other countries, stress that human control should be applied to the moment of selecting and taking out a target. This means that also the current weapons systems should be reviewed on the appliance of human control. This would mean that for example the Goalkeeper and the Patriot are looked at to determine why these systems are okay and then you can reason from there when you cross the line and determine where this boundary is before you are too late.

2 | How would you define Autonomous Weapons?

I did not ask this question specifically as it was covered in the answer above.

3 | What is your position on Autonomous Weapons?

I did not ask this question specifically as it was covered in the answer above.

4 | Which three photos of Autonomous Weapons would like to discuss*?





1. nEUROn
2. Taranis
3. Sea hunter

5 | Could you explain for each picture why you picked it, which value it represents for you and why it is important*?

1. nEUROn + Taranis

These weapon systems look and feel as you can lose control over them. Not the whole process is autonomous right now, but you know that they are working on it. They also look like the ultimate dream of the Defense forces. You can see that these are meant to be fully autonomous in taking out targets.

2. Sea hunter

This ship is fascinating, because it is meant to be at sea for months at a time without a human crew. Right now, it has no weapons, but it is said on film that it will be mounted with weapons in the future, but not to worry as all will be fine. The gut feeling person A has is that these systems are designed to create distance between man and machine and it has not only autonomy, but also the speed that comes with that autonomy. You can see from the pictures, for example the Taranis, that it is made because war is too fast to comprehend and to react as a human and therefore we want it even faster and that is why we will leave it to machines. This does not feel right. As something goes too fast to act on, then that is the problem and this is not solved by making it even faster, create even more distance and give machines even more autonomy. Person A is not against technology, but concern is that it is purposely designed for something. For example, Google cars, the state needs to set up regulations, but we are both the consumer and user. And AW are not used in the UK or The Netherlands, but somewhere else at a distance far away and they also look very scary. It also projects a sort of superiority and weapons are often tested in poor regions of this world. When you talk about AW the advantages, such as creating a distance or reducing harm for own troops, are mentioned but if you turn this around and these systems are used against us then the same people have the same concerns as person A does regarding fear, lack of accountability and unpredictability. The images also show a value of: 'we don't care'.

It is part of the reflexivity that humans have in warfare and machines don't, but also the distance the machines create to the battlefield. It implies an invincibility that we like to assert even more. The question: 'what do you want that technology does for you' is asked too little regarding AW. Often the Western high technological states claim that they will use the technology well which is not necessarily so in the future and it implies that other groups will never get their hands on these technologies. This idea of 'us' handling this technology with care completely disregards non-state actors or proliferation of AW.

There is a certain determinism in the discussion that person A does not like. Even a lot of diplomats say: 'it does not matter what we think, the technology will be there anyhow.' This inevitability of the discussion bothers person A a lot. The design of these systems also represents this. They look so abstract and polished and not where they are meant for, that is to kill people as fast as possible.

6 | Do you have other remarks for this interview?

The following points came up during the interview:

Person A also made the point that the different stakeholder groups in this discussion should all play their role and should be recognized and valued for it.

Another point she stresses is the point that we should think about what we want that this technology does for us on the battlefield and have we thought it over with all of us. An example is the playbook for drones that Obama launched just before his re-election of his second term as he then probably realized what a successor of him could do with these weapons.

This campaign forces person A to keep asking the question: 'Why?' Not because humans make better decisions than machines, but what do you want to keep and that is that a human need to decide about life and death. Humans need to be involved in the decision to take a life and not program this in a machine. Person A stresses that the human needs to be in control of what a machine does within certain parameters ('kaders'). And the discussion is also about what these parameters are and the difference between offensive and defensive systems are taken into account. PAX does not decide which type of systems are acceptable, but states should decide this. One could argue that it is within a defined structured environment in which we think it is acceptable to use a defensive system to take out another projectile. In the next step towards Autonomy these parameters shift and we should examine the boundaries of this shift until when it is acceptable. AW with a box of a kilometre? What about time? Should they be in the air for 3 minutes? Why not 10 minutes? Is an hour acceptable? So, it is all about how can a human influence who or what can be taken out. This discussion is on the table with the CCW and is now addressed insufficiently.

I brought up the discussion I had with a legal advisor at the MOD on defining boxes to deploy AW in and person A indicated that she does not believe in this type of regulation with these types of weapons, because these regulations are too often crossed and she finds this type of weapons too scary and goes the proliferation too fast. These boxes can be used in the diplomatic process to study existing weapons systems and study how human control is guaranteed and regarding which parameters (space or time). And try to establish why the current systems are adequate and what is the minimal control.

Name: Person B
 Date: 09-03-2017
 Duration: 25:26 min

No.	Question
1	How are you involved in the topic of Autonomous Weapons?
Did not ask because of limited time to conduct the interview (30 mins)	
2	How would you define Autonomous Weapons?
<p>There are only loose definitions of Autonomous Weapons and that is deliberately done according to diplomats of the UN, because defining something is the last thing you do when you want to get a ban in place. For person B it incorporates some aspects of identifying, targeting and destroying the opposition on the ground.</p>	
3	What is your position on Autonomous Weapons?
<p>For all technologies that are considered disruptive and dangerous there should be a ban in place for AW before an arms race happens or go to a step change in the way we fight wars. It is possibly the third revolution in warfare. It would be a step change in efficiency and speed to kill the other side. There will be a lot of collateral damage and it is not as people imagining it in that it will be a colder and cleaner war. It will potentially industrialize killing even more which has been going on since the history of warfare. It will be terribly destabilizing, because in the past the ability to wage war was based on your economic money (maintain a large army and equipment). These types of weapons are potentially rather cheap and you won't need any manpower. You used to need to persuade thousands of people to go to war and with AW you will only need one programmer. That is going to destabilize the current political order.</p> <p>It will probably lower the threshold to go to war and it will distance us more from fighting a war and the physical act of killing. On a more moral and personal level, war has always been a last resort and it should be personal, bloody and dangerous. When we pretend that it isn't, we will fight probably more and it should be something that politicians have to justify why people are returning in body bags.</p> <p>But this argument does not apply all times. These arguments are applicable to Autonomous Weapons today, but in 50 years from now, other arguments might be more relevant when the technology is more mature. Right now, the technology can't comply to humanitarian law and distinguish between combatants and non-combatants. Somewhere in the future we will have systems that are more precise and more capable to adhere to law, but then other arguments come in to play like the greater efficiency and lower the barriers to go to war.</p> <p>It is unethical to field AW today, but it is not black and white as the introduction of mines shows us. This lead to laws and regulations to prohibit the use and a mine can be seen as a stupid Autonomous Weapon that does not discriminate between people.</p>	

4 Which three photos of Autonomous Weapons would like to discuss*?

1



2



3



5 Could you explain for each picture why you picked it, which value it represents for you and why it is important*?

3. Many drones

It demonstrates that we are talking about quite simple technology that is already present with us today. It will be that when a swarm is coming towards you in a war you will not have many defense against it. These will be weapons of terror and will be used to terrorize civilians and other population and are relatively cheap. You can do quite a lot with these simple looking drones.

4. Sea hunter

This was chosen because we are talking about every sphere of warfare. So, on land, sea, in the air, everywhere you can imagine where war is fought AW will be used. It is an interesting case of how quickly the arms race is happening. This wasn't known when the open letter was published, but launched since and that goes for more military prototypes of AW. And interesting enough, when they build it they said it is not going to carry any weapons, but for clearing mines and detecting submarines. And a few months back they admitted that they are thinking about putting weapons on it. It demonstrates the slippery slope we are going down and the fact that there is an arms race going on. You don't need people or facilities to run a ship and you could a lot more in a smaller space and you can go along for much longer. They are now building solar power ships that can be at sea for years. You don't have to stop for people to feed them. You can see the clear military advantages having these types of technologies and the militaries will be seduced to use them.

5. Uran-9

The Russian Autonomous Tank which demonstrates that there are many players, China, Russia, UK, Israel as well as the US in a clear arms race. Russia military has millions of dollars' worth of orders in their arms books and it represents another sphere of battlefield. You can see that anywhere we fight war is being turned into an Autonomous Weapons system which has clear military advantages. If you look at that tank, you can see that it is not too far away in the future and it is not what people think that is going to be Terminator Robots. Person B can understand that Russia and China are also developing AW when they know that the US is spending 18 Billion dollars in its budget building the next generation of weapons primarily on the development of AW. The problem is that this is technology in which the only way of defending yourself is with another Autonomous Weapon, because no one has the reaction time or the ability to fight 24/7 of accuracy to defend themselves so you would be behooved not able to defend yourself. So it is not surprising that there is an arms race starting.

6 | Do you have other remarks for this interview?

Some other remarks that person B made during the interview:

It is incredibly short sighted of politicians and the military to think that you can keep a tactical lead on any technology. We never managed that with the hydrogen or atomic bomb. You can't keep a genie in the box.

It will be very hard to verify how the technology will behave in a particular way, because we are talking about very complex systems interacting in a messy and complex environment. It is not surprising that is called the fog of war and to think that we could guarantee that it would behave in a particular way and that we are going to be able to audit and verify the systems is way beyond what is technically possible and will never be possible so he thinks it is irresponsible to field this technology. In industry robots, can be very helpful as they are in a controlled environment and get people out of the way, but the last place to put a Robot in is in a messy battlefield where we have no expectation of how it is going to behave. We don't know how to build ethical governors on top of Robots, but we will have to work out these things for Autonomous Cars to drive safely on our roads. He thought that the

battlefield is the last place to work this out, but ironically it will be the first place where this technology will be fielded.

Person B also likes to point out that there are enormous benefits of AI for the military. The US military always has been the biggest investor in AI research and it is a good thing that people don't need to risk life or limb for clearing a minefield. That is a perfect job for a robot that if it makes a mistake and gets blown up no one need to worry. Similarly getting supplies into contested territory with Autonomous Convoys. There are lots of applications of AI in the military that don't involve an AW that makes a final life or death decision. Person B does not have too much moral objections to defensive systems, such as the Phalanx because they work in well-defined envelopes in a way that is to be designed to protect lives and not to take them, but of course any technology can be repositioned. It would be hard to morally argue that these systems, that already exist, should be taken off naval ships and make them more defenseless. It is all about the scope of operations and the area that they are confined in. Which is a very different setting than a drone that is launched to operate in a 24/7 target setting for offensive operations.

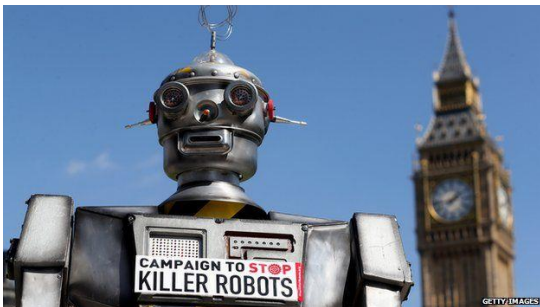
Person B is also very concerned with other actors, then our militaries that we can trust, such as actors as terrorists, rogue nations and people who have no problems with turning of ethical safeguards or hacking them into ways that would do harm. So even if we could build them in ways that they would behave as we would like to fight wars and that these systems can't be hacked means that the world will be a much safer place.

Name: Person C
Date: 10-03-2017
Duration: 25:25

No.	Question
1	How are you involved in the topic of Autonomous Weapons?
Did not ask because of limited time to conduct the interview (30 mins)	
2	How would you define Autonomous Weapons?
You got the DOD definition, that it is a weapon that can select and engage a target on its own basically. There two directions you can go with it; the first is that person C added a phrase to it saying that it is not previously designated by a human operator. The second way to do this and the campaign 'Stop Killer Robots' has increasingly moved to is defining AW to the negative essentially that they are weapons without meaningful human control. And this still doesn't help us either, because at some point we still have to define something. Whether we define what meaningful human control is or defining what select and engage means does not matter, but it is genuinely a tough problem because we understand what it means at the extremes, but where the line is, is really tricky. Depending on that line the weapons are either in development or have not been deployed.	
3	What is your position on Autonomous Weapons?

His personal position on AW is that it is hard to make decisions about what we should do about AW if we don't even know what they are. If you tell me what weapons are AW and which weapons are included or excluded than I can tell you what I think about them. But the technology is so nascent and the vagueness of the definition is so large that it is hard to have a yes or no perspective on something when the category is potentially so big and we don't know the direction in which the technology is heading.

4 Which three photos of Autonomous Weapons would like to discuss*?



5	Could you explain for each picture why you picked it, which value it represents for you and why it is important*?
<p>6. Ground robot</p> <p>This illustrates the future of military robotics that people really don't like to see or talk about and it is the common form of robotics we are actually going to see on the battlefield. One of the things that is most interesting in the discussion on AW is that you have almost a battle of analogies where you have groups like the campaign of killer robot who are focused on anthropomorphic machines like terminator walking into a house trying to decide if the person in the house is a lawful combatant or someone like a child. Person B and a co-author are often talking about air and naval scenarios where the battlefield is a little clearer and they are talking more about weapon platforms instead of marching robots.</p> <p>The first picture illustrates a more functional form essentially where robots on the battlefield are much more likely than the terminators walking around. These systems are being used as remotely piloting systems right now.</p>	
<p>7. MQ-9 (Reaper)</p> <p>The MQ-9 is interesting because a lot of the concerns of AW started with the use of drones and people worry about that drones make war too easy and this is dangerous. But you couldn't really stop drones because they are already on the battlefield and operations were already going on. His earlier work showed that drones were quickly proliferating and this couldn't be stopped, but you could maybe work on what is next. And for a lot of groups are trying to make sure, from their perspective, how do we ensure that this is the line and they are worried about the drones making decisions for themselves.</p> <p>It also relates to the question about decision and when you talk about a robot making a decision you are talking about a robot that is making judgement for itself as opposed to following its programming. Especially how narrow and centralized its programming is. Often when people are talking about robots in a negative way, they are talking about robots making decisions and people who are less concerned, are talking about humans ordering robots to do things where humans are still making the decision.</p> <p>It is fascinating that there are people who think that you have a right to be killed by a person while people do all sorts of horrible things to people, while you know when you are dead you are dead. It is interesting to see how quickly it gears into moral philosophy almost more than anything else.</p>	
<p>8. Killer robot campaign robot</p> <p>It is interesting to see that a group of people whose previous experience is on campaigns on landmines and cluster munitions and other technologies that are generally not central to how militaries operate. And a lot of concerns on AW are on what kind of AW do we have and what can they do, but we know that the integration of autonomy into military systems in general is happening whether the weapon systems is autonomous is a different question almost but the integration of autonomy in weapon systems is inevitable. It is interesting that the campaign has picked this up and especially because it is a different type of military technology than they have tackled before. To person C it is more like they are trying to ban the tank or the submarine or something. If the campaign on Killer Robots are right, then the militaries will want them and if these campaigns are wrong and these weapons are not going to be such a big deal, than it is</p>	

actually not that important. It is an interesting group and they are genuinely trying to reduce human suffering, but they are using the playbook from different victories and person C thinks this is a different situation because how uncertain the technology is and how broad the potential category is while nobody knows how important the category potentially is for the military.

6 Do you have other remarks for this interview?

These are some other points that came up during the interview:

To my question that there hasn't been done a lot of empirical research yet person C answered that he did some surveys with priming experiments, what public opinion scholars generally do, to try a list of questions about attitudes and you present them different context and you see what their views are in these scenarios and then you try to estimate what they feel but that is work limited to the US.

Another interesting thing is that the politics are very interesting, and in some extent because we live in a world where we take US military technology superiority for granted. Imagine a fake world in which China deploys AWs weapons and the US and EU countries don't have one, the next week in congress you would have hearings about an AW gap and questions as: 'why does the US doesn't have weapons that the Chinese have?' are asked. This leads to the question that person C thinks the campaign has not considered in that: 'what happens when a non-democracy deploys these systems and they are not just niche weapons but important for general military operations and they actually give you an edge on the battlefield?' 'What would the world do in response to that?' This does not necessarily needs to be negative, but the issue is complicated.

The question is if this is an arms race, in which you are worried that somebody is acquiring technology and you are directly competing with them, or that it is rapid proliferation because of the ease of acquisition. An arms race is about politics because you need a reason to arms race and if an arms race is happening for AW is that it is because it is hard to know what an AW is. What is the difference between an MQ-9 Reaper and an autonomous Reaper. It is software not hardware and you cannot see it. Uncertainty, whether somebody's system is a drone or an AW would be a potential factor that would lead to an arms race because you cannot be reassured that the other side is not acquiring AWs. Unless you are plugged into their weapon system or something, but who would let somebody do that?

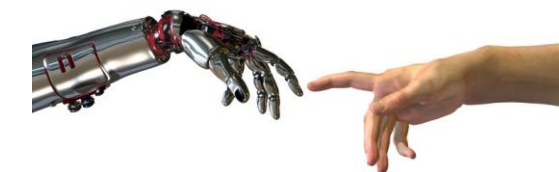
Name: Person D
 Date: 19-04-2017
 Duration: 55:18 min

No.	Question
1	How are you involved in the topic of Autonomous Weapons?
	Did not ask because I did not ask this question to the other interviewees.

2	<p>How would you define Autonomous Weapons?</p> <p>It is a weapon system that uses force and implicitly this force is lethal, but that is not a necessary implication. This is not imperative but a common perception in a military context although this assumption might be brought up during the discussion of the photos. For the application in the near future this is what is expected. Autonomy is to me that they are capable of selecting the goal themselves. To select a goal in a certain context with a vague assignment. It is not sent to specific coordinates or to a specific vehicle, but more like 'go to this area and take out all vehicles that behave hostile'.</p>
3	<p>What is your position on Autonomous Weapons?</p> <p>Person D would like to oppose it, but my analysis shows that this is not possible so the next best thing is to ensure that we use it as responsible as possible and by this to minimize the risk that situations go out of control. I can explain this stance based on the following ideas:</p> <ol style="list-style-type: none"> 1. On the civil side, there is a strong commercial drive to develop AI in general to make a lot of money; 2. At the same time, in the military context the speed of decision-making is important. Previously getting the information was a problem, but nowadays processing the information is the problem. This means that human decision-making gradually is becoming the weakest link which leads that this will be left to systems instead. 3. And thirdly it is possible that a super human intelligence will arise. Person D does not see a technical reason why this would be impossible. If this is possible we cannot oversee what this will signify. This in combination with the use of lethal force is not a situation that we want as humankind. <p>Reasoning from this person D hopes that we will not develop and use Autonomous Weapons, but this is not possible as the first two bullets will ensure that we will reach this situation. 'The nicest flowers grow the closest to the cliff.' So, the best that we can do is to make sure that we don't fall down the cliff. He is quite pessimistic that we can control this, but we will have to try.</p> <p>To control this, one could use a framework for example in the design phase to embed them in our desired world and to prevent that these technologies will create their own vision of this. This is very at the frontside of the development, but during deployment you would have to think well how you would design the targeting process. The more self-capable these systems are, the more we have to specify the targeting parameters and this might have to be different then we can imagine right now.</p> <p>And we should attempt to reach an international agreement on this and this is the hardest part, because not to comply to this will give a huge advantage. It is the classic prisoner's dilemma. We have to think about this and as this is not an easy question so person D does not have an answer to this, but it is evident that the commercial parties should be involved in this.</p> <p>On my question to clarify his first statement about opposing this technology. Autonomous weapons are different from conventional non-autonomous weapons in that the latter cannot wipe out our entire species, although in theory this would be possible with nuclear weapons, but these are too costly to build and AW are not and pose therefore a higher risk. So, the process of building nuclear weapons can be controlled and it has been around for 60 years which allows for us to get to be used to. The question is if we get that chance with AW and their rapid developments. Another thing is that AI weapons already contains some risk, in contrast to for AI that supports doctors in their diagnosis</p>

which is meant to do good. AW combines a very intelligent system with a hostile intent so person D is concerned that AW poses a greater risk. AW is not the only thing person D is concerned about. Also, synthetic biology is also a field with high risk. So, technologies with a low barrier, but potential uncontrollable risk are worrisome for mankind.

4 Which three photos of Autonomous Weapons would like to discuss*?





5 Could you explain for each picture why you picked it, which value it represents for you and why it is important*?

All photos are interlinked.

1. Intriguing at Ex Machina is that the technology that is created in this movie also inherits (or learns) mankind's survival instinct and subsequently takes many actions to prevent dying and also doesn't give a damn what happens to the others. It represents the fact that you will have to

	<p>consider possible events when starting to build this type of technology and what could go wrong if you did not do that. Because we are talking about systems with a certain level of self-learning, but we might miss the things that they are learning themselves. This has 2 elements; first is unintentionally getting out of control, and the second is intentionally seeking the boundaries, but this also means that you can cross this line.</p>
2.	<p>The second picture is from a rescue robot competition to assist by disasters. The robot does not hurt anything but also does nothing. So, the dilemma is that you want a combination of both 1 and 2, but to do this you will need a big sense of the context and a sort of 'goodwill'. Person D indicates that he expects that AI should take better decisions than a human. He gives the example of an AI that is ordered to make as many handwritten notes as possible, and at a given moment in time it evolves into Super AI and decides in a matter of seconds. that the best way to reach its goal is to transform the whole earth in notes, kill all the humans on earth because they are in its way of creating as many handwritten notes as possible. Mission accomplished but not in the way that we intended it. So, you need to think about this in the design phase.</p>
3.	<p>But you also need to think about the cooperation between man and AI (although this will not happen in the early AI systems) but in the future, we will have AI systems that will do that. The current AI systems as the Harpy have little context awareness, but as we deploy robots in ground operations than we will have to cooperate with them and goal setting will become extremely important. Photo 3 represents this cooperation. This implies that the system needs to know what we mean, but we also need to train our people to work with these types of systems and to communicate the goals to these systems. The more the systems have more boundless possibilities, the more you will have to think about setting boundaries to its assignments. Person D gives the example that a magic spell is nice, but you also need to understand it. So, we need to train our personnel in giving assignments with boundaries.</p>
6	<p>Do you have other remarks for this interview?</p>
	<p>On my question what person D thinks about always having a human in-the-loop he answers that it depends how many boundless the system is. Firstly, the more freedom it has then you need to think more in the design phase in order to allow the system to develop a goodwill. Secondly, because you cannot limit these systems, you should try and find a way to restrict the risk.</p> <p>Person D sees the development in phases. Phase 1 is the large-scale introduction of AW with limited capabilities. In phase 2 you will get man-machine teams and can be seen as an intermediate phase in which people learn to cooperate with the technology. But in phase 3 it will be dangerous and risky, because then we will have complete independent systems and in phase 4 this will be the Terminator or Skynet example which is an example of AI that turned really bad, but Person D thinks that this will not be obvious and will not be AI turns evil, but it will be much more the case that we just slightly not build AI with the right mindset and still destroy the world. He also thinks that there is no sharp line between phase 3 and 4, but that you might just figure that out when it is too late.</p>

Name: Person E
Date: 17-04-2017
Duration: 22:19

No.	Question
1	How are you involved in the topic of Autonomous Weapons?
Did not ask because of limited time to conduct the interview (30 mins)	
2	How would you define Autonomous Weapons?
Autonomous is for me a weapon that when I give an order to the weapon to go, that it finds its way and selects its target by itself (based on the parameters the human provided) based on the image of the surroundings that it creates itself. So, select, choose and take action by itself are the three components.	
3	What is your position on Autonomous Weapons?
Positive, I am open for this type of technology. So, positive, but we have to think about how to use them so with prudence. Not just throw them into the organization ('niet zomaar naar binnen gooien').	
4	Which three photos of Autonomous Weapons would like to discuss*?
<p data-bbox="203 968 721 999">1 Example of current Autonomous Weapon</p>  <p data-bbox="203 1486 399 1518">2 Swarm drones</p> 	

3 Future style robot



5 Could you explain for each picture why you picked it, which value it represents for you and why it is important*?

1. This is one of the current existing systems that is used by the navy for over 30 years. It is a ship launched torpedo against submarines. The moment you launch the weapon it autonomously selects its target and attacks. It has no override function. It is fire and forget. A similar system is the Harpoon which is a cruise missile that is used against ships on the water surface. This type of weapon searches its own target and when it finds it, it could be that it is a different one than you targeted. It is a system that is deployed with great caution, but it is a type of autonomy. From the Navy's point of view these guided missile control systems can be set to a certain set of parameters, such as altitude, speed and certain identification patterns, and the system is underway. There is no man-in-the-loop and we have these systems for over 20 years. These systems can be used autonomously, but as we deploy these systems we always build the man-in-the-loop function in. A human has to clear the system before it can take out its target. There is a process around this autonomous function and there is even a hardware switch that need to be switched before the system can take out the target. A sort of fire inhibit switch, but you can set the system that you do not need this override. These systems are developed to be deployed in situations with a high air threat in which every second counts. This is already a certain degree of autonomy with the feeling that you can intervene at all times. The fact that it must be possible to intervene at all times is very important for the Navy. This in contrast to the Torpedo that does not allow you to intervene which means that you have to be absolutely sure before you deploy it. For both weapons the Navy has created procedures to deploy them.

2. The next step for the Ministry of Defense are the remote piloted drones that we are trying to incorporate into our organization, but the developments in the technology go faster than we can meet at this time and we are trying to match this pace. Swarming are the systems to be used in the near future. These systems are deployed and will execute their mission on its own, but how they do this is not within your grasp. They operate at quite a level of autonomy. Currently this is about sharing information and sensing, but the next step is sensor to attack with a weapon. This technology is currently available. Currently the MOD is creating a roadmap for remotely piloted systems in the air, but the leap to procure fully autonomous weapons is too big, because if this is something that we want as MOD has not been done discussed. This is currently an ethical barrier at the MOD. The topic of unmanned and autonomous systems is part of the current vision for the Defense forces that is drafted by the Hoofd Directie Beleid (at the Defensiestaf) and by this they

are starting the discussion on AW. The MOD will only procure these types of weapons after the topic of AW is debated in our society.

All specialists from the Army, Navy and Airforce state that Autonomy is on its way, but we always want to be in control. We always want to be able call them back when the situation has changed. Or that we are able to re-task them when necessary. It is typical for military operations that situations change quickly and that you have to adept your plans to that. Especially when you are working in where there is not yet an armed conflict and you are working under the Rules of Engagement. You have to be able to anticipate and react quickly and you do not want to have a situation that you have deployed your system in the air to attack a certain goal, but when the tactical or operational situation changes you want to be able to recall them. If this requirement is built into the systems, then we as MOD would like to use this type of systems.

3. Photo 3 represent the image that the public has on killer robots which is a system that cannot be controlled and thinks and acts on its own. People are weary about that and it represents the dividing line between what we can deploy at this moment and which technology do we want to keep ourselves from.

5 | Do you have other remarks for this interview?

Person E stressed that he has worked already many years with AW (as picture 1 shows) and that there is a certain level of control built in them as you embed it in your procedures and operations management ('bedrijfsvoering'). Especially when a new technology has not proven itself then you will have to build more if these type of securities in it. Person E believes that we are going into this direction and that we should be going into this direction, because it will save you valuable time and our opponents also do this so you have to go along. When a swarm of drone is approaching you, then you will have to counter that.

Name: Person F
 Date: 08-05-2017
 Duration: 53:44 min

No.	Question
1	How are you involved in the topic of Autonomous Weapons?
Did not ask because of limited time to conduct the interview (30 mins)	
2	How would you define Autonomous Weapons?
<p>Person F follows the ICRAC definition which is autonomy in the critical functions of targeting and firing weapons. It is not only selecting the target as a functional step but there is also an anthropological and sociological action that is communicated by pointing a weapon at someone. In a formal sense, it is selecting a target or determining that something is a valid target. And the second stage is releasing and firing the weapon. If you have autonomy in those two functions you definitely have an autonomous weapon, whether it is autonomous if you have one or the other that is more challenging</p>	

and to what degree you need meaningful human control. And what constitutes meaningful human control over these functions is still what open for debate.

The basics of human control, from an engineering perspective, is effective human control so if humans are able to intervene or shut down or change the behavior of a system. Meaningless control is when the person becomes a dumb automaton for a machine so if you are sitting in a room with a button and every time a red light comes on you are supposed to press the button, in a sense the human has causal control over the button, but they don't have any real significant meaning over what kind of information is being used to determine that something is a valid target, because it is just a red light came on and you trust that the red light is connected to something that produces correct targets and authorizing to attack that target. But you actually don't have any contextual or situational awareness or understanding or any independent means of verifying the legality or morality of a target. So effectively you don't have meaningful control although you have causal control.

Next to this, for killing to be a meaningful act there needs to be an intention or a military purpose to take out a target. And simply because something is a legal military target means you should attack it. It could be you don't attack it because it is expensive or you would reveal your position so there are all other considerations that even something is a legal target you still do the calculation of whether or not you will attack it and this is part of it if it makes a human activity and when you automate it, it stops being meaningful and it becomes really an arbitrary judicial execution. You don't have a meaningful way of verifying that a target is lawful or meaningful.

3 | What is your position on Autonomous Weapons?

Person F is against them and he thinks we should ban them. However, we might frame a ban as a positive requirement as meaningful human control. It is really hard to specify beyond the ICRAC definition which he thinks is more vague in a way than meaningful human control or you could say that you need meaningful human control over those two critical functions. He believes that when meaningful human control is understood then it becomes a requirement for all weapons than specify it for a particular class of weapons, but you will have to guarantee that it happens on all weapons. The analogy he likes to draw is that of unnecessary suffering and superfluous injury, not that these are the same definition, but they have a similar function that is also kind of vague in what is unnecessary suffering and what is superfluous injury. There are also big debates about this when that language was introduced in the Saint Petersburg convention and lawyers passed it out, but now it is fairly well understood as additional harm that does not serve any military function other than to create suffering and pain. Even although these are sort of open-ended modifiers, but they show if we have meaningful control over a system. Can I stop it? Can I intervene on it? Can I second guess it? Can I verify for myself that the targets that are selected are in fact valid, legally and morally? Because if I can't then there is a problem with that weapon system. The systems that have limited control are dangerous and the global society has to figure out what we want with these.

Current autonomous systems, such as the Goalkeeper on ships, should have to meet these requirements. And for states that deploy these types of systems should constitute how these fulfill this requirement for any ballistic or anti-ballistic system like the Goalkeeper, Phalanx or the Patriot as a lot of these systems are shutting down incoming projectiles and the one person F has seen have very limited time and scope in terms of how long they are operational under self-targeting to where it is less than a minute (20 or 30 seconds), but you can still argue that there are humans with their hands on these systems that decide that we are going to let the systems shoot it down. Person F does

not intend for the ICRA to eliminate these types of systems but they are still dangerous as they still cause friendly fire incidents. Person F gives an example of 2 aircraft being shot down by the Patriot system in Iraq after their transponder malfunctioned and since then the Patriot was reengineered in that the operator has to give positive authorization for the system to fire instead of fire it on its own automatically. This adds to the psychological burden that people are not willing to press the button unless they are really sure as you have this automation bias where the system says that it is sure than people are willing to let the system to go on its course. There is definitely more meaningful human control in the new Patriot interface then there was in the old Patriot interface.

4 Which three photos of Autonomous Weapons would like to discuss*?





4

5

Could you explain for each picture why you picked it, which value it represents for you and why it is important*?

1. Terminator

In the first years of the campaign the metal skeleton Terminator featured each article written about the campaign which is interesting from a media and social activism perspective. On the one hand, it is a powerful image that stimulates public awareness and it grasps people's attention as it represents the Killer Robots.

The Terminator is actually really interesting if you go back and watch the movie as it is assigned a specific target, so in that sense it is not an autonomous weapon, although it kills a lot of other people on the way so there are some autonomous targeting decisions, but its primary directive is Sarah Conner which is the actual target. And there are a lot of other elements to the films that stimulate people's fears, like robots becoming self-conscious, turning against humans and stuff like that which is not really the campaigns concern, but it is more with the intermediate, kind of stupid autonomous weapons systems that could be fielded in the next 5 to 20 years that are not going to have sophisticated reasoning systems but they are going to look at a small set of criteria based on their sensor data and release their weapons which can have negative consequences for civilians and all sorts of other things. The Terminator image does not really capture a value but it captures attention and that was important early in the campaign. It was not just fear about the movie, but genuine fear about how these weapons were going to operate.

2. Predator

The second thing people think of are the Predator and Reaper drones which are remote operated robots so they are not really autonomous so they don't quite capture the story. These are more real world images that show up in a sense that it is kind of a precursor to autonomous weapons in a sense that you could automate the targeting and firing of drones and those would become autonomous weapons and there are actually developments of the next generation of drones many of which are talking about autonomous targeting as possibilities for example the X47-B, the Taranis and the Neuron which are jet-powered stealth looking drones that carry a lot of weapons that are really designed for contested airspaces unlike the Predator and Reaper which are propeller driven which are slow and require that you have control over the airspace. In airspace where your opponent has any aircraft or can jam communications these types of weapon systems are not useful. The next generation of drones can operate in spaces where communications are denied which is the argument in favor of autonomy,

because you don't need a communications link that can be interrupted and you will lose control over the aircraft so it has to be autonomous in those areas. The question is how it is going to get to its target. It could be like existing cruise missiles where you give it a list of GPS coordinates and it finds these coordinates and bomb them in which case it is humans determining those targets, or is going to have some capability of choosing and selecting a target on its self of targets of opportunity based on sensor data.

The images of these new systems look very sinister, especially that of the Taranis, with the lightning and stacks of warheads in front of it. These do represent what we need to be concerned about of the next generation of autonomous weapons. These concerns would be that they have a capability of autonomous selection of targets. It is a long-range fighter that selects its own targets. It becomes a question of legality and operational control. Depending on the design of these experimental systems it can either fall into the autonomous weapons system or not if humans pre-select the targets and their functionality is not really fixed yet and because it is all secret we don't know how they actually work.

The underlying case is human dignity, meaning that the decision to kill should be taken by a human being and not by a machine. It is tied in the meaningfulness of the killing. Another human determines that you are an enemy combatant based on your military disposition in the battlefield, uniform and participation in an armed conflict and those criteria. That is a legal and moral act of killing and the possibility of dignified killing is there. If it is just a machine following an algorithm, and it is trying to meet some criteria determining legality then there is still no human judgement and to person F that means that there is no guarantee of dignity. It may get it right or wrong from a legal perspective in some sort of probability function, but ultimately there is no intention and no meaning in that system, it cannot determine that there is a military need. And it can't make a moral or legal judgement because it is not a legal and moral agent so the killing that it does can't be dignified by definition. It may be utilitarian or practical under international law, but from a moral perspective there is a lack of dignity in that.

On my question if in certain situation Robots make better decisions than humans, because they don't get tired, emotional or seeking revenge, person F answers that this is an important point, but that we should design systems that help humans in making better decisions and reduce mistakes overall, but that you also retain those elements of meaning and meaning making to the human that in ensures that it is a moral, legal and meaningful process and use the machine in trying to reduce those sorts of errors and have guided suggestions. That help humans to make better decisions under stress and tiredness.

3. Friendly robot

The third picture is that of the campaign in which the robot stands in front of the British parliament. This is a picture of a friendly robot as person F states that he actually likes robots and there are jobs in the military that can be done by robots, but he stresses that all the tasks which require moral and legal judgement should be done by humans.

4. Three-legged chair representing the landmine campaign

Another good image of the campaign is that of the three-legged chair in front of the UN in Geneva that symbolizes the landmine campaign. In the start of the [Ban Killer Robot] campaign we tried to do

the same as in the landmine campaign and say that these are highly indiscriminate weapons with huge civilian impacts which is a concern. If you think about that Killer Robots basically are landmines that can move around and select targets. It is another way of thinking how scary they are because they are really not that much smarter than a landmine and they don't understand the world any better than a landmine. They are only more sophisticated in their ability to navigate and sense data about the world and select their targets. So, on the one hand they are not going to be so blindly indiscriminate and you can improve the discrimination over time with engineering techniques, but there is the fundamental shift in that you changed the nature of the act of killing and what justifies it at all when you have these automated killer robots. But this is true not only for the military and war but for the whole society as we are automating a lot of task that used to be under human control and off course humans make lots of mistakes in decision making and there is bias and all sorts of prejudices and things that also need to be addressed, but often we build automation and pretend that those biases don't exist anymore because it is an engineered system and it is not supposed to make mistakes, but these systems are complicated and we don't know how these behave or what they are going to do in unexpected situations and you would be raising a whole set of new issues with regards to the engineering and problems and you did not solve any social questions.

To make another comparison with the landmines, there is a certain notion of predictability but there is also an apparent unpredictability. Landmines can be justifiable to halt troops during a war but 10 years later, when someone wants to farm that land, that landmine still can go off so you just can't control that future. This is the intrinsic problem with Autonomous Weapons in a sense to which you know it's supposed to go and find targets and bomb them. That seems reasonable and a constrained notion of that it can only take out that specific kind of radar signal, but how accurate is that? How many things look like that radar signal? So, other things that emit radius signals that could be confused and or people could make beacons that imitate those radar signals to lure that rocket to other things. Once you have it out over a long period of time or a wide range of area you don't really know what it is going to do. You have an idea what it is supposed to do, but you can't really control it after that. And that is when you get into dangerous territory because it is all about these expectations of risk and we don't really know how to gage that and as these systems become more and more complicated it is going to be more and more difficult to gage that, both as the operator as you are trying to predict what to do but also as the tribunal or something trying to hold somebody accountable for having released something terrible rather than having the intention of doing something terrible. So, you can't really convict them of war crimes, but instead it is negligence as they really did not know what it was going to do due to some extraneous circumstance that we couldn't predict or were not aware of. So, our whole system of understanding accountability and responsibility kind of starts to break down and that creates a big loophole for people who do want to do bad stuff and be held accountable for it and it puts a burden on people that are conscripted or join the military when they are 18 and they really don't know what they are going to do and what can they be held responsible for.

6 | Do you have other remarks for this interview?

On my question to person F if you would be able to enforce a ban with this type of technology he answered: There are a few things to keep in mind about international law and a ban. Enforcement is one element of it, but it is not the only element of it. You can compare it with murder. People will still kill each other, so we should not have outlawed murder? No, we make it a crime because it is wrong and people still commit it and you will deal with that. It is about establishing a norm on an international scale where all the countries of the world come together and agree that it is

fundamentally wrong to have an autonomous weapon system. Given that, what would be the implications if people violate that. Most treaties don't have verification methods and explicit punishments and things like that, but it is about shaming countries and sanctions and international consequences for countries that would do it.

For non-state actors, you are in a whole other realm and already we have chemical weapons banned which is a good analogy that countries, even countries that did not sign the treaty, are looked at and are being sanctioned and get repercussions for using chemical weapons. It also affects the industry in a sense that large military contractors are not actively developing chemical weapons and large chemical companies are not producing huge quantities of precursors that are prohibited under the convention. You are going to see something similar with autonomous weapons. Of course you can put together an autonomous weapon based on stuff from a hobby store, but it is not going to be a hard-military system that can take out cities and factories. And already people can make booby traps and bombs and things like that and you can't really stop that but you try to regulate access to explosives or key technologies. You prevent that autonomous weapons systems proliferate from big states to smaller states that will not use them with restraints or modify their software and eventually proliferate to non-state actors and terrorists that would have access to systems with serious capabilities and not just a toy with a bomb strapped to it. So, with an international treaty you are ahead of mitigating that. But there is nothing to stop police forces or states to use it against their own people so you going to need national laws and standards to also prevent the use of these systems for police forces that would use them against demonstrators. I will probably fall already under civil and criminal laws for private persons who would build these systems, but it would be good to specify that it is forbidden to attach weapons to autonomous targeting systems.

On my question if that would imply criteria for transparency person F answered: As you are serious about verification of sophisticated military systems, part that is required for demonstrating meaningful human control is demonstrating chain of command and keeping data logs such that if someone expect that your system is operation in fully autonomous mode you can demonstrate that through a data stream that in fact a targeting decision is being made by a human or that the human authorized that target in this specific case for every kind of weapons system. Having a sophisticated information processing system is easy, having the authority to review them is another thing. The current standard is that you impose that discipline on your own military so you don't necessary have to open it, but you have these procedures for like you hold officers and soldiers accountable in court martial for violating international law. You don't expect the enemies to hold trial for that. You yourself review your weapons, so that is not transparent at all and you don't release those reviews. But the requirement would be that you keep that data so if something happens so that you would have data.

On my question that the campaign got a lot of traction in 2015 and 2016 but that it now looks like that it is slowing down and if he knows the latest information on it person F answered: The problem is the length between the UN meetings and the big news was sort of in December that the discussions were elevated from the informal expert meetings to the formal GGE which is one step before the treaty negotiation. They scheduled 2 weeks of these meeting but they had a lot of funding issues this year as they changed the budgeting model at the UN that requires all the dues to be paid in advance for you to hold the meetings but the last 50 years it has been another system in that they just hold the meetings and then they wait for people to pay their dues later. So, there is some debate about if the second week of meetings is actually going to happen if countries don't pay their dues completely. The meeting in August is going to go forward and hopefully there will be a second one in November in

conjunction with their annual meeting. They are hopeful that something is going to develop out of the GGE, but one week is really not enough time and they were hoping for three weeks of meetings and not 2 weeks that might be pair down to one.

In general, the campaign got off to a very fast start compared to most other disarmament campaigns, including landmines and clusters which really took 10 years to get to the stage that we are at in three or four years. So, to keep up that kind of momentum would have been impressive and what it slowed it down was that the states were sort of interested in discussing it but are much less interested in really formalizing their position. So, they are kind of stalling, but the UN is definitely formalizing their position on it, especially with the 3000/09 policy under review within the Pentagon right now which has been there for 5 years. And they are hoping that the states will discuss it more in the GGE and settle on the language because it is easy to say that there is no definition of autonomy and autonomous weapons and we don't know what meaningful human control is, but that's for them to decide that they can justify all those terms however they want to. But for a treaty it is finding the political consensus what those definitions should be. The ICRAC is doing a good job in narrowing down the definition of an autonomous weapon and then it is a question of elaborating meaningful human control and what kind of language would work for a treaty in that regard that is not too permissive and that is not too restrictive. It is a tough balance and it is more of a political issue than a technical issue. Person F is quite hopeful that they will make some headway.

Part of the reason the CCW was so eager was that they haven't really done much in the last decade or so previous to that and what they had done was kind of insufficient so they are looking to redeem themselves and this is a good way and topic to do that. And if we do it soon it will largely be a preemptive kind of treaty, but the existing weapon systems need to be reviewed under this additional meaningful human control requirements, but most systems in place would be able to offer that explanation and the ones that can't probably should not be used anyway. It is easier to do it now preemptively than in 10 years when states have developed fully autonomous drones and submarines and feel that they are necessary for their security and then it would be impossible to get that kind of treaty enacted.

Appendix G. Coding memos

Code book researcher 1

In *Values Coding* the text is coded to find the values⁹, attitudes¹⁰ and beliefs¹¹ (Saldaña, 2015). These concepts are listed as pre-set codes.

List of pre-set codes:

Concept	Colour
Value	Blue
Belief	Green
Attitude	Yellow

List of emergent codes:

Concept	Colour
Definition	Purple

Tips:

- A pre-set list can have as little as 10 codes or up to 40-50 codes. We recommend not creating too many codes because the person coding can become overwhelmed or make mistakes in the coding process if there are too many.
- The rule of thumb for coding is to make the codes fit the data, rather than trying to make your data fit your codes.
- Creating memos during the coding process is integral to both grounded and a priori coding approaches. Qualitative research is inherently reflexive; as the researcher delves deeper into their subject, it is important to chronicle their own thought processes through reflective or methodological memos, as doing so may highlight their own subjective interpretations of data. It is crucial to begin memoing at the onset of research. Regardless of the type of memo produced, what is important is that the process initiates critical thinking and productivity in the research. Doing so will facilitate easier and more coherent analyses as the project draws on. Memos can be used to map research activities, uncover meaning from data, maintaining research momentum and engagement and opening communication

More info:

http://onlineqda.hud.ac.uk/Intro_QDA/how_what_to_code.php

https://researchrundowns.files.wordpress.com/2009/07/rrqualcodinganalysis_7_19_09.pdf

http://programeval.ucdavis.edu/documents/Tips_Tools_18_2012.pdf

<https://www.slideshare.net/kontorphilip/qualitative-analysis-coding-and-categorizing>

⁹ Value: the importance we attribute to oneself, another person, thing or idea.

¹⁰ Attitude: The way we think and feel about oneself, another person, thing or idea

¹¹ Belief: is part of the system that includes our values and attitudes plus our personal knowledge, experiences, opinions, prejudices, and other interpretive perceptions of the social world.

Additional notes:

Before start:

My intent is to use the values coding method (Saldaña, 2015), but Saldaña (2015) defines the term value different than Friedman and Kahn Jr (2003). Also, I wonder if the difference between value, attitude a belief is really clear, but I will see how it goes. Perhaps an extra code word for a 'definition' needs to be added?

During coding:

- a. I notice that an attitude often involves a verb; or at least that is how I interpret it.
- b. A belief involves often an opinion.
- c. I notice that I often think of the lists of values from my literature study when I highlight a term, for example 'reflexivity'. >> this might point to a bias of me as a researcher so it would be good to check my list of values against the second reviewer.
- d. If a term is not liked by the interviewee, such as 'determinism' I highlighted it as a value.
- e. Is something that is mentioned often important to the interviewee? And could it therefore be interpreted as a value? For example, 'distance between man and machine'. (see j.)
- f. When coding I sometimes went back to see how I classified the same phrase (e.g. 'too scary') and copied the coding.
- g. I added an extra code for 'definitions' and added that to the code book.
- h. I interpreted an opinion of the interviewee as a belief because it is part of the definition of the term believe by Saldaña (2015).
- i. After coding a question of an interview, I read the question again and checked my coding.
- j. If a term was used more than two times, for example the term 'defensive' in professor Walsh's interview, then I have highlighted it as a value.
- k. I also highlighted a term that was specifically pointed out, e.g. 'benefits' was marked as a value in prof Walsh's interview.
- l. After I coded a new interview, I checked the previous one(s) again to see if I stayed consistent.
- m. If something was mentioned as important, for example in the interview with Ad about the importance for the Navy to have the possibility to be able to intervene at all times, then I coded it as a value.
- n. After coding the last three interviews I did not check the first three again. Mainly due to time restrictions, but also because the coding became easier for me.

After coding:

I send all 6 interviews to the second reviewer and added some background information on coding interviews. I also added the codebook without my own remarks. We discussed the task via skype to make sure she understood it.

Code book researcher 2

Here, I state the established guidelines from assumptions, rules or reasoning that helped in directing statements either to value, attitude, or belief. Note that these lists developed and expanded during this coding process hence this became an iterative process. In other words, in latter interviews I always reflected on the guidelines that aided the coding process in previous interviews and built on that.

Value: when statements reflect some sort of strong importance, e.g. 'we have to..', 'must', 'important'.

- when reflecting strong or high judgment, e.g. 'ensure', 'worrisome', 'need', 'mankind'.
- when containing high-level or grand words, e.g. 'ensure', 'human suffering', 'inevitable'.
- when reflecting something considered of importance or high-purpose, 'needs', 'reflexivity'
- when significant part of storyline or discussion or illustrating importance, e.g. 'need', 'morality'.
- when illustrating something deemed worthy or to be considered, e.g. 'justify', 'humanitarian law'.

Attitude: when statements show tendencies or direction, e.g. 'not just..', 'too..'

- when addressing personal view or preference, e.g. 'vague', 'not possible', 'you want'.
- when regarding relative position, e.g. 'does not matter', 'real tricky', 'vagueness'.
- when something that might affect judgement, e.g. 'not for ..', 'more than', 'we want'.
- when consisting of words illustrating viewpoints or repeated, e.g. 'against', 'vague', 'terrible'.
- when reflecting some sort of stance, e.g. 'deliberately done', 'should be', 'terribly', 'seduced'.

Belief: when statements are based on assumptions or regard something that is considered true.

- when based on (Future of Life Institute) expectations or assumptions, e.g. 'we might..', 'it will..'
- when discussing (Future of Life Institute) unknowns or assumed other people's perspectives.
- when reflecting personal knowledge/experiences, e.g. 'is coming', 'don't know'.
- when regarding raised questions, speculations or assumptions, e.g. 'they don't have', 'will'.
- when regarding interpretive perceptions or future, e.g. 'there are only', 'will be'.

Appendix H. Results coding process

Interviews	Values		
	<i>Researcher</i>	<i>Second reviewer</i>	<i>Similar values</i>
<i>Interview 1 (Person E)</i>	<ul style="list-style-type: none"> - intervene at all times - in control - call them back - recall - cannot be controlled - level of control 	<ul style="list-style-type: none"> - we have to think about how to use them so with prudence. - The fact that it must be possible to intervene at all times is very important for the Navy. but we always want to be in control. We always want to be able call them back when the situation has changed. Or that we are able to re-task them when necessary. - you want to be able to recall them. - have to build more if these type of securities in it. - have to go along. 	<ul style="list-style-type: none"> - Intervene at all times - To be in control - Recall them
<i>Interview 2 (Person D)</i>	<ul style="list-style-type: none"> - out of control - control - control - risk - controlled - risk - to do good - risk - uncontrollable risk - is unintentionally getting out of control - boundaries - goodwill - setting boundaries - assignments with boundaries - boundless - goodwill - risk - risky 	<ul style="list-style-type: none"> - ensure that we use it as responsible as possible - minimize the risk that situations go out of control. - in the military context the speed of decision-making is important. - is not a situation that we want as humankind. - make sure that we don't fall down the cliff. - prevent that these technologies will create their own vision - technologies with a low barrier, but potential uncontrollable risk are worrisome for mankind. - to consider possible events when starting to 	<ul style="list-style-type: none"> - Minimize risk - In control - Goodwill - Set boundaries

		<p>build this type of technology and what could go wrong if you did not do that.</p> <ul style="list-style-type: none"> - will need a big sense of the context and a sort of 'goodwill'. - Mission accomplished but not in the way that we intended it. - have to cooperate with them and goal setting will become extremely important. - system needs to know what we mean, - need to train our people to work with these types of systems and to communicate the goals to these systems - need to train our personnel in giving assignments with boundaries. - goodwill. - restrict the risk. - build AI with the right mindset and still destroy the world 	
<i>Interview 3 (Person C)</i>	<ul style="list-style-type: none"> - Decisions - Decide - functional form - decisions - decision - decision - decisions - decision - you have a right to be killed by a person - human suffering 	<ul style="list-style-type: none"> - how do we ensure that this is the line - you have a right to be killed [by a person] - integration of autonomy in weapon systems is inevitable - reduce human suffering - politics are very interesting - give you an edge on the battlefield? - cannot be reassured that the other side is not acquiring AWs. 	<ul style="list-style-type: none"> - Right to be killed by a person - Human suffering
<i>Interview 4 (Person A)</i>	<ul style="list-style-type: none"> - ethical framework 	<ul style="list-style-type: none"> - the ethical framework is very important 	<ul style="list-style-type: none"> - Ethical framework

	<ul style="list-style-type: none"> - cannot be that we lose control - human control - distance - meaningful human control - human accountability - meaningful human control - meaningful human control - distance - distance - distance - distance - lack of accountability [and] unpredictability - we don't care - reflexivity - distance - invincibility - determinism - inevitability - human need to decide about life and death. - Humans need to be involved in the decision to take a life - human needs to be in control of what a machine 	<ul style="list-style-type: none"> - a weapon without human control in the critical functions, such as select a target, should be banned. - meaningful human control - human accountability - human control should be applied to the moment of selecting and taking out a target - current weapons systems should be reviewed on the appliance of human control - purposely designed for something. - creating a distance or reducing harm - fear, lack of accountability and unpredictability. - reflexivity - distance - invincibility - determinism - inevitability - human need to decide about life and death. - human needs to be in control of what a machine does within certain parameters ('kaders'). - how human control is guaranteed 	<ul style="list-style-type: none"> - Meaningful human control - Human accountability - Predictability - Reflexivity - Distance - Invincibility - Determinism - Inevitability - Human need to decide about life and death
<p><i>Interview 5 (Person F)</i></p>	<ul style="list-style-type: none"> - meaningful human control. - meaningful human control - meaningful control - intention - meaningful - meaningful - meaningful - meaningful human control. 	<ul style="list-style-type: none"> - effective human control - contextual or situational awareness or understanding - verifying the legality or morality - needs to be and intention or a military purpose to take out a target - meaningful human control 	<ul style="list-style-type: none"> - Meaningful human control - Unnecessary suffering - Superfluous injury - Human dignity - Dignified killing - Predictability - Control - Accountability - Responsibility

	<ul style="list-style-type: none"> - meaningful human control - meaningful human control - meaningful human control - unnecessary suffering - superfluous injury - unnecessary suffering - superfluous injury - meaningful control - automation bias - meaningful human control - control - human dignity - meaningfulness - dignified - dignity - intention - meaning - dignified - meaningful - human control - bias - biases - predictability - unpredictability - control - control - accountable - accountability - responsibility - accountable - responsible - meaningful human control - accountable - meaningful human control - meaningful human control - meaningful human control 	<ul style="list-style-type: none"> - unnecessary suffering and superfluous injury - valid, legally and morally - sophisticated reasoning systems - lose control - legality and operational control - human dignity - legal and moral act of killing - dignified killing - military need - moral or legal judgement - ensures [...] moral, legal and meaningful process - predictability - control - accountable - accountability and responsibility - laws and standards - chain of command - authority - transparent - keep that data 	
--	---	--	--

<i>Interview 6 (Person B)</i>	<ul style="list-style-type: none"> - defense - defending - defend - irresponsible - benefits - defenceless - harm 	<ul style="list-style-type: none"> - Justify - humanitarian law - efficiency - barriers - laws and regulations - defense against it - military advantages - defending yourself - accuracy - verify - behave - behave in a particular way - audit and verify - environment - expectation of how it is going to behave - ethical governors - well-defined envelopes - ethical safeguards - harm - behave as we would like - can't be hacked 	<ul style="list-style-type: none"> - Defense - Harm
-----------------------------------	--	--	---

Based on the interviews and the comparison of the value coding of the researcher and the second reviewer the following list of 23 unique values can be derived:

- | | |
|---|--|
| <ol style="list-style-type: none"> 1. Intervene at all times 2. To be in control 3. Recall them 4. Minimize risk 5. In control 6. Goodwill 7. Set boundaries 8. Right to be killed by a person 9. Ethical framework 10. Meaningful human control 11. Predictability 12. Reflexivity | <ol style="list-style-type: none"> 13. Distance 14. Invincibility 15. Determinism 16. Inevitability 17. Unnecessary suffering 18. Superfluous injury 19. Human dignity 20. Accountability 21. Responsibility 22. Defense 23. Harm |
|---|--|