Ethical Dynamics of Autonomous Weapon Systems

Marcus Frølich
Department of Science and Industrial Systems
University of South-Eastern Norway
Kongsberg, Norway
e-mail: 218867@student.usn.no

Mo Mansouri
Department of Science and Industrial Systems
University of South-Eastern Norway
Kongsberg, Norway
e-mail: mo.mansouri@usn.no

***Abstract*—Introducing an Autonomous Weapon System (AWS) to warfighting has a multitude of intertwined military and civilian consequences. As we are in the early phase of this shift towards autonomy, we can still influence development and legislation. That makes such investigations valuable. Much has been written and debated about the ethical implications of AWS, but little about how these ethical questions interact. This paper models the ethical dynamics of AWS, using a causal loop diagram that shows causal links between topics related to military use and civilian acceptance. Using this novel approach to describe the ethical implications of introducing AWS, otherwise hard-to-find interactions are highlighted. Based on this model, two leverage points and their effects are discussed: feedback on soldier infractions and live/recorded battlefield data.**

***Index Terms*—autonomous weapon system, ethical dynamics, causal loop diagram, systems thinking, systems engineering**

I. Introduction

An Autonomous Weapon System (AWS) is a system that can select and engage targets without further intervention by a human operator [1]. They represent both the present and future of modern warfare. With them comes both advantages and a wide range of new challenges, especially concerning the ethics of war. It is the strong belief of the authors that AWS will increase in variety, autonomy, and capability, without wanting to take a position on whether this is ethically, morally, or legally correct or not.

This paper seeks to highlight some of the most relevant advantages and challenges with AWS to have a more well-informed discussion, comprising a wide range of contexts for the systems. It will also highlight some leverage points and solutions where potential downsides can be limited. However, this paper is not meant to be an exhaustive list of all arguments that illuminate every side of the discussion.

When the term AWS is used throughout this paper, there is never a particular weapon system in mind, but the broader sense of all physical autonomous weapon systems. This can be autonomous drones, autonomous patrolling vehicles, stationary mounted weapon stations able to autonomously detect and engage targets, etc. It does, however, not include the cyberspace domain.

To spark a well-informed discussion on the ethical dynamics, this paper is structured as follows: Section II describes the ’human and the loop’ terms with some discussion on challenges with them, Section III gives an introduction to how AWS is entering the battlespace, Section IV gives a brief overview of where the future of AWS is heading, Section V describes categories and questions of legality for AWS, Section VI introduces a causal loop diagram to enable discussions on the dynamics for AWS acceptance and use, while Section VII highlights some leverage points, Section VIII discusses some future research topics, and finally, Section IX concludes the paper.

II. Human and the Loop

Some of the most frequently used terms when discussing AWS are the terms relating to where the human is in the loop. Following are descriptions of how the authors see these terms relate to AWS and some of the challenges in using them.

Human-in-the-loop

When robots can select targets and deliver force only with human command, we say the human is *in* the loop. Removing the human would limit the functionality of the robot.

Some argue that the term is very ambiguous as the human will always be somewhere in the loop, without necessarily specifying where [2]. For this reason, the authors argue to save the term for systems where the main actions will not be performed unless it has human input. One example is when an autonomous system can find and track a target, but the action of ’pulling the trigger’ is left to the human operator. Such border control systems in South-Korea and Israel are later discussed.

Since systems in this category require human input, they are often called supervised or in some cases semi-autonomous, but can not be called truly autonomous, hence not covered by the term *autonomous* weapon system. Semi-autonomous weapon systems can, however, also be that a human operator selects a target that is engaged autonomously [1].

Human-out-of-the-loop

In the opposite case, the human is said to be *out* of the loop when the system can perform all its tasks without human input.

For an AWS, this means it can autonomously both select and engage a target without any human interaction, like the later described air defense systems that can autonomously detect and engage incoming air targets. This is full autonomy and is often referred to as an unsupervised system.

 ISBN: 978-1-61208-838-9

Human-on-the-loop

The middle ground is when the human is *on* the loop, meaning the system can operate autonomously under the oversight of a human operator who can override its actions.

An issue with this term that can make it considered de-facto 'out-of-the-loop' is when a human is given only a fraction of a second to make a veto decision [3] or the system lacks adequate or sufficient human supervision [4].

At the same time, human soldiers already have to make snap judgments in the field under highly demanding constraints, and often with limited situational awareness [3].

A risk when overseeing autonomous systems is the 'automation bias', the tendency to trust automated systems over own judgments, even when provided evidence that the system is unreliable or wrong [5]. We might attribute more capabilities to an AWS than it truly has [2].

III. Entry of Autonomous Weapon Systems

Depending on the cultural background, 'killer robots' like from the movie Terminator [6] might be the association of an AWS. The first appearance of the term 'robot' was in 'R.U.R. (Rossum's Universal Robots)' by Czech writer Karel Capek in 1921 [7]. Since then, the idea of highly intelligent robots taking over the world has inspired popular culture. Autonomous systems capable of applying lethal force are already a reality, albeit not in the dystopian sense some are envisioning.

Current systems include the Israeli 'Iron Dome' system at the Gaza border, 'CRAM (Counter Rocket Artillery Mortar)' in Baghdad [7], and the Norwegian made 'NASAM' system deployed by, amongst other places, the White House in Washington D.C. These are all air defense systems capable of autonomously taking down incoming air targets.

Other systems are targeted at humans. Samsung SGR-A1 is a South-Korean system deployed at the demilitarized zone towards North-Korea, able to recognize human shapes and command them to stop and surrender [4]. Israel also has a system at the Gaza border with automated kill zones [8]. Both systems are currently depending on a human in the loop to apply lethal force, but reports confirm that in both cases they have the capability to deliver lethal force without human input [8].

A highly successful system is the teleoperated system used in Afghanistan by both the American Army and its allies. These are weapon systems that are operated by a human soldier, but have additional capabilities, such as keeping the aim of the weapon stably on the target found by the operator, even if the system is mounted on a moving vehicle. Some see these systems as an extension of the soldier, thus not autonomous by themselves [9]. It is easy to envision a future where human input is no longer needed.

The now-retired American remotely piloted aircraft Predator sparked discussions about modern warfare where the operator no longer need to be physically present at the place where the action takes place. As Peter Singer, military expert and author, stated in an interview with E&E: "[...] 20 minutes after being 'at war' you're sitting at the dinner table talking to your kids about their homework" [7]. The successor of this system is Reaper, a remotely piloted aircraft that also includes autonomous capabilities.

IV. Future of Autonomous Weapon Systems

The U.S. Army Research Office has founded research into an algorithm that can be used to rank the most valuable targets in a terrorist network [10]. One can foresee a system that connects this algorithm with the Reaper drone to automate a drone strike if the algorithm decides that killing is more ethical than capturing the target [3].

For now, the human remains in the loop, but there is almost inevitable that humans in many cases will be omitted from the loop. As discussed, this might even be the case for systems where humans are designed to be *on* the loop. According to Peter Singer in 2009 [7], Pentagon already back then had a research project called 'Taking Man Out of the Loop'. In 2007 there was a proposal for research by a US Army division stating that "[...] Fully autonomous engagement without human intervention should also be considered [...]" [8].

Some even predict a future where the humans are no longer fighting the wars, but withdraw from them completely and let the robots shoot it out [3][11].

V. Just War Theory

Just War Theory is the tradition and justification of how and why wars are fought [12]. To ensure a morally justifiable war, multiple criteria must be met. It is common to divide the evaluation into three groups: 'Jus ad bellum' (going to war), 'Jus in bello' (fighting a war), and 'Jus post bellum' (after a war). The three core principles of Just War Theory are discrimination, proportionality, and military necessity [4]. Discrimination means distinguishing between enemy combatants and noncombatants (civilians). Proportionality means that the harm is in balance with the gains of the action. Military necessity means that the end goal of the war is achieved through the least amount of harm.

Note that in this paper the term 'Just War Theory' will be used in the broadest sense to cover all aspects of a just war, including International Humanitarian Law, Rules of Engagement, and the Geneva and Hague Conventions.

Since concepts like autonomy, artificial intelligence, and robots are all general terms with no clear-cut definitions, there will likely not be one simple legal assessment we can all agree upon. Even if a country concludes that its development and use of AWS is in accordance with Just War Theory, others might argue against it.

In the aftermath of a war, historians can generally agree on who was the attacker and defender. During the war, both sides will always declare they have an ethically justifiable cause to enter the war. When evaluating tactics during a war, the two sides will also disagree on, e.g., what are mistakes and what are deliberate actions. Although both attack and defense can be just, deploying AWS for defensive strategies is generally easier justifiable.

Regardless of this, humanity is best served with systems that answer to Just War Theory to the biggest possible extent. That is why it is important to discuss compliance in more contexts than purely to define the legality of a particular AWS.

Jus ad bellum

In the case of justification for going to war, there is generally one major concern with AWS that is highlighted: lowering the threshold of entry to war. AWS generally means less political risk. This lowers the barriers in entering conflict without enough forethought and without exhausting nonviolent options, thus contradicting 'Jus ad bellum' and potentially making it unethical [2][8][9][13]. Some even argue that 'risk-free war' might put the civilian population at increased risk from terrorist attacks at home and abroad, by making terrorism the only way to fight back [14].

A reason why AWS is seen to give less political risk is what is often referred to as the 'Dover test' [9]. It has its name from the Dover Air Force Base in Delaware, the base where soldiers are returned from the front line in flag-draped coffins. The 'test' determines how much war casualties affect the electoral chances of the sitting political administration. This has been a major inhibitor of military action by the US since the Vietnam War [9].

The argument of AWS lowering the threshold of entry to war is typical for any significant technological advance in weaponry and tactics [8]. Some argue this makes it not worth discussing, especially when evaluating the legality of AWS. The authors believe it is nevertheless important to discuss how to create a system that raises this threshold, as it has both legal and ethical implications.

Jus in bello

It is a principle under 'Jus in bello' that someone can be held responsible for deaths and infractions that occur in the course of war [8][9]. A human soldier can behave unethically or make errors, but can be held accountable for it. For an AWS is it more unclear, as neither the system's designer, developer, maintainer, nor the military officer who deployed the AWS had any intent to cause a crime [4]. It would be unfair, and hence unjust, to hold, e.g. the commanding officer of the AWS entirely responsible for actions over which he or she had no control, both to the commander and any resulting causalities [4][8]. This lack of clear responsibility for possible war crimes is often referred to as the 'accountability gap' [4], and can make AWS unethical.

As Lin is quoted in Defence Robotics [2]: "If a human orders a robot to storm a hideout and shoot the insurgents inside, but the robot detects mostly children and women, what should it do? [...] with potentially greater situational awareness, a robot could have reason to refuse". The principle of responsibility includes the consequences when obeying orders that are known to be immoral [8]. At the same time, do human soldiers have sufficient information about the situation to determine if the order is morally correct? The capability of an AWS to process large amounts of data might give it an advantage over human soldiers in evaluating the greater moral implications of an order [11].

Some see the problem of discrimination as the most difficult aspect of AWS [8]. Failing this key principle can be seen as a reason alone to ban such systems. Current pattern-recognition technologies can discriminate between civilians and uniform-wearing soldiers based on images [4]. This might lead the enemy to cease wearing uniforms, thus increasing the risk for civilians. The authors believe that the problem of discrimination is a technological issue that will eventually be solved for AWS, at least to a level that outperforms humans. In the meantime, some argue a ban is correct, equivalent to the ban on antipersonnel landmines that does not take into account a hypothetical future improvement of the equipment [4][8].

A middle ground might be found that avoids a complete ban, while the technological development and discussions on discrimination are still not settled. An AWS could be made to only identify and target weapons and weapon systems, not the individual(s) manning them ("target the bow or arrow, not the archer" [8]). By disallowing AWS to select and engage humans as targets, the limit is at their capability of initiating a kill order [1][2][8].

Jus post bellum

Establishing truce and lasting peace should be the goal of any war. This requires that the different parties see the other as a serious partner, despite the differences. A potential issue with the introduction of AWS is if it leads to a 'moral deskilling' [3], later discussed in Section VII-A. This can make it harder to secure meaningful peace.

VI. Causal Loop Diagram

The dynamics of how the general public sees AWS and whether it is accepted or not is complex. To help comprehend the various aspects that contribute to the varying use and acceptance, a causal loop diagram is created in Figure 1. The development of the diagram is based on literature research.

An arrow indicates interrelation between nodes. When two nodes change in the same direction, the causal link is noted positive ('+'). The opposite direction is noted negative ('-'). Closed cycles in the diagram can be either positive reinforcing ('R') or balanced ('B'). Delays are noted with crossed-out links. Nodes that are colored were found in the literature to be the most important nodes. The two green arrows will be discussed in further detail in Section VII.

When evaluating if a causal link is positive or negative, the context is important. One can compare military strategy, tactics, and operation. What is found as a positive link for e.g. strategy, might contribute negatively at the tactical level.

The diagram is simplified by putting the use and acceptance of AWS together, based on an assumption that they are proportional and directly intertwined. It is the use of AWS that the general public accepts, and the politicians approve

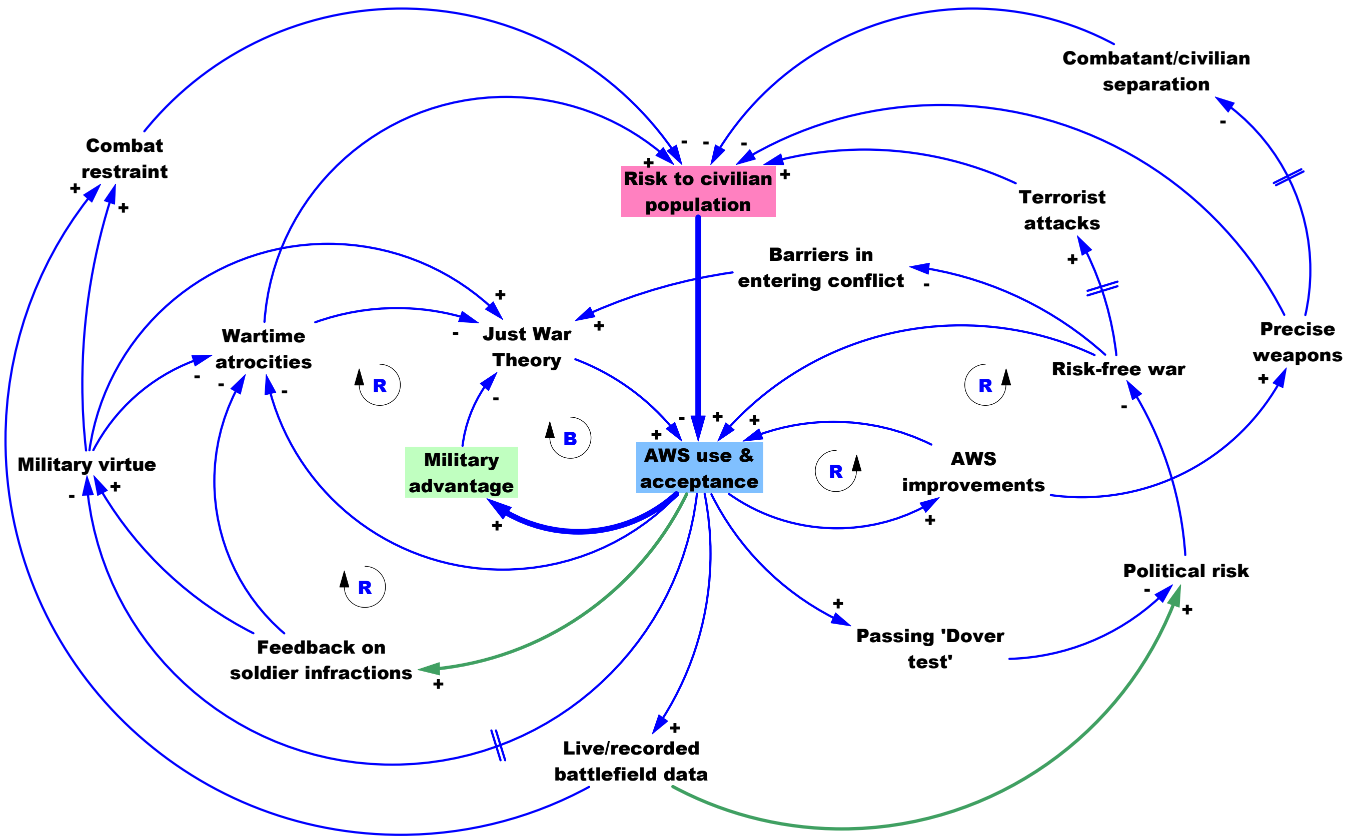

Fig. 1. Causal loop diagram, military use and civilian acceptance for AWS.

the development and use based on acceptance in the general public.

To model the dynamics, the most important nodes and causal links are included based on their contribution to the ethical dynamics. In order to ease discussions by ensuring better agreement on the meaning of the nodes, they are briefly described below. Most of them are discussed in further details in other sections of this paper.

Military advantage

The main reason for the military to push for the development and deployment of AWS is the military advantage they bring to the battlefield.

Risk to civilian population

The main concern with the introduction of AWS is the potential risk it may have to the civilian population, on both sides of the war.

Feedback on soldier infractions

An AWS may objectively assess and report soldier infractions from the battlefield.

Live/recorded battlefield data

An AWS may increase the amount and quality of data received from the battlefield, e.g. video streams.

Military virtue

Military virtue such as courage, integrity, honor, and compassion is deeply embedded in the human soldiers, but may both be reduced in humans soldiers that are removed from the battlefield and hard to embed into AWS.

Combat restrain

Combat restraint ensures the least amount of force necessary is used to reach a goal.

Wartime atrocities

A grim fact of wars is the occasional atrocities, an important node to keep an eye on.

Just War Theory

Used in this paper as a collective term for all laws, regulations, and ethical guidelines governing a just war.

Passing 'Dover test'

Refers to the number of war causalities and how this affects the dynamics.

Political risk

The political decisions of using military force come with a risk of upsetting the general public, i.e. voters.

 ISBN: 978-1-61208-838-9

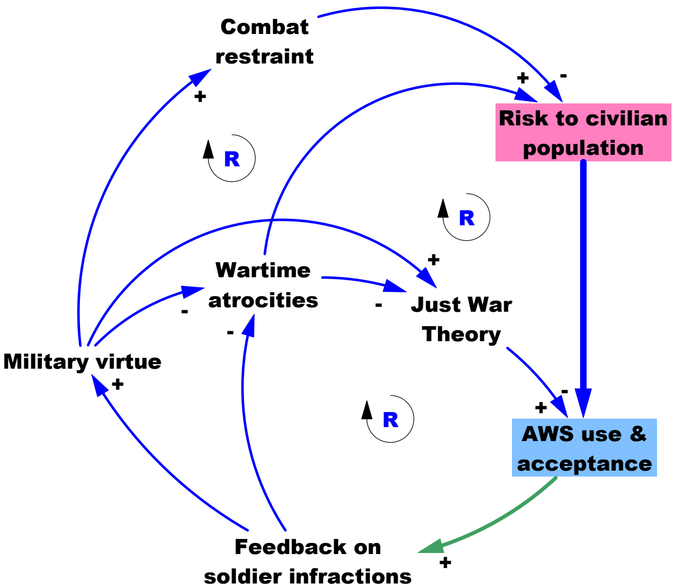

Fig. 2. Causal loop diagram, focus on ’Feedback on soldier infractions’.

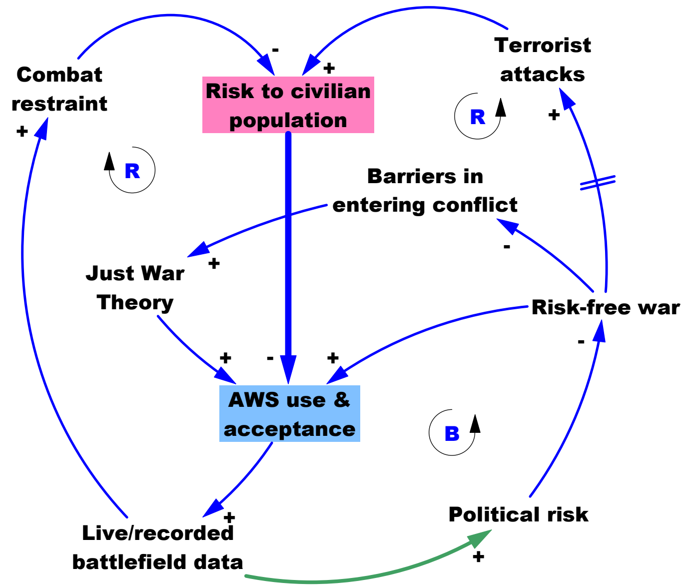

Fig. 3. Causal loop diagram, focus on ’Live/recorded battlefield data’.

Risk-free war

The concept of going to war without having to fear the consequences.

Barriers in entering conflict

There are always arguments for and against going to war. The threshold for when a war decision is taken is important.

Terrorist attacks

An opponent may resort to terror as part of warfighting, both abroad and at home.

AWS improvements

As experience is gained and general technology advances, the AWS will become improved.

Precise weapons

A property of AWS that is generally seen as superior to humans, is its precision on the battlefield.

Combatant/civilian separation

The (visually) clear differences between a combatant and a civilian.

VII. Leverage Points

An important reason to model the dynamics of a system with a causal loop diagram is to identify potential leverage points [15]. These are points where a small shift can have a big impact. Following are two of the leverage points found when analyzing the causal loop diagram in Figure 1.

A. *Feedback on Soldier Infractions*

The positive causal link to the node ’Feedback on soldier infractions’ seems to be a key link. All the consequences from this link are isolated in Figure 2. It symbolizes a chance to use the AWS for objective and unbiased evaluations concerning the alignment of soldiers’ habits and decision patterns with norms of military honor, courage, and restraint [3][8].

This increase in feedback may lead to reduced soldier infractions and wartime atrocities, as there is a constant awareness that such events will be reported [8].

Military virtue, such as courage, integrity, honor, and compassion is crucial to help fight a just war [3]. When each part in a war sees the other as a professional actor with honor and a moral purpose, it can motivate restraint. It also helps the soldiers to keep their moral connection with society. During the war’s aftermath, military virtue helps secure meaningful peace between the parties. Having feedback on soldier infractions will help to keep the focus on, and maintain the credibility of, both the human soldier’s and AWS’s military virtue.

There are already suggestions for architectures with ethical governor, ethical behavioral control, and ethical adaptor and a responsibility advisor [8]. There is also research into ’Artificial moral intelligence’ [3]. These can become tools that help AWS evaluate ethics and morale.

A great concern of Vallor [3] is that the introduction of AWS will have a negative impact on military virtue. If moral decisions during a war are performed by pre-programmed AWSs and not cultivated by humans in the chain of command, it can lead to a ’moral deskilling’ of the military. To maintain such cultivation and expertise on moral virtue, repeated and habitual practice is essential.

As the extract in Figure 2 shows, all the loops created by the ’feedback’ link are positive reinforcing loops. Thus, from the arguments presented in this paper, an increase in feedback

 ISBN: 978-1-61208-838-9

will increase the use and acceptance of AWS. At the same time, when seen isolated, it has no negative effects on the other nodes leading back to ’AWS use & acceptance’.

One could argue that by knowing that an introduction of AWS will increase feedback on soldier infractions, this will in and of itself increase the acceptance of AWS from the general public. This will create an additional positive feedback loop. At the same time, soldiers might have a negative attitude towards a system constantly evaluating them.

B. Live/Recorded Battlefield Data

AWSs rely on a constant stream of data to take actions on the battlefield. Some of this data is processed internally on the AWS, while some will be transferred back to an operator. For instance, a drone will usually return a video stream, whether it operates fully autonomous or with a human operator or supervisor. An advantage of all the gathered data is that AWSs can rapidly share this and help improve decision-making [11].

Much like the potential for a dedicated system giving feedback on soldier infractions, knowing that the AWS will live stream and record battlefield data will restrain human operators, AWS developers, and chiefs of command from deploying fully autonomous systems. There already exists procedures to record data from weapon systems and to review this internally in the case of special battlefield incidents.

As depicted in Figure 3, the big leverage point lies in ensuring that this data is not only kept within the ranks of the military, but becomes publicly available. This way, it increases the political risk of using AWS in an unethical manner.

With an increase in political risk, it creates a balanced loop that helps avoid an uncontrolled escalation of unethical AWS use. The importance of this can be seen by studying the Vietnam War, often referred to as the ”first televised war” [16]. During the middle of the war, the number of reporters grew tremendously, bringing all of the war’s brutality home to the living rooms of the American public. Although we have become more custom to it, live and recorded video and other data from the battlefield will likely lead to an increased aversion to war by the general public [8].

No military general is likely to allow live streaming of all data from war, as this can reveal military secrets and give up military-strategic advantages. Journalists already have ”freedom of information” laws that guarantee access to government documents. Similar rights of access by journalists could be expanded. The technicalities of how such data can be shared are no further discussed herein.

An example of the importance and ramification of sharing video footage is the July 12th, 2007 Baghdad airstrike, where two US Army helicopters launched air-to-ground attacks on a group of people, including civilians and reporters. The crew was heard laughing at some of the causalities. This video was not shared by the military, but leaked to the whistleblower website WikiLeaks by former US Army soldier Chelsea Manning [17].

VIII. Further Research

Expanding the causal loop diagram may make it harder to use as a discussion tool, but could help find more leverage points. For example, by finding more ways to influence the barriers in entering conflict and combat restraint, or to discover further negative causal links which must be offset.

Qualitative analysis of the causal loop diagram is given. Further detailing of the model can also enable quantitative evaluations. Computer simulations can, for example, be useful both to quantify and visualize how an effect on one node or link changes the whole dynamics of the model.

For a war with AWS to be just, according to the principle of ’Jus in bello’, someone needs to be accountable for the actions of the AWS. What role this person has and how distributed the responsibility is may influence the ethical dynamics.

If a restriction is set on AWS to not target humans, how will this alter the dynamics? Instead of only analyzing how the flow in the causal loop diagram change, it might make more sense to develop separate diagrams for different levels of restrictions on AWS.

Further research should be done on decision support systems with prediction models that can judge when, how, and under what constraints it is ethical to deploy AWS. Some research into tools for evaluating ethics and morale is already mentioned [3][8].

Section II categorized AWSs into human-in-the-loop, human-out-of-the-loop, and human-on-the-loop. When analyzing ethical issues, it would be beneficial to have a much deeper classification tree.

Whether a side of the war is seen as the attacker or defender, aggressor or retaliator, will greatly affect the evaluation of ethics. For AWS, an autonomous border patrol system deployed on one’s homeland is likely to have more acceptance than, e.g., a drone flying over foreign soil. Adding this dimension of context to the ethical dynamics of AWS will likely be useful.

Ethical dynamics that change depending on the context are important to clarify. For example, by investigating the three abstraction levels of military strategy, tactics, and operation. It should be investigated how a change in viewpoint might affect the causal loop.

IX. Conclusion

In this paper, the authors pursued to model the ethical dynamics of autonomous weapon systems, and identify leverage points and mitigations to dampen some of the challenges.

Analyzing the proposed model proved successful in finding positive contributors to AWS acceptance and maintained ethics. One suggestion was to introduce and embed systems to give feedback on soldier infractions, and another was to find ways to ensure that the public gets access to some of the data collected by the AWS on the battlefield. The latter was demonstrated to introduce both a positive balancing force into the dynamics and reinforce some desired effects.

REFERENCES

[1] United States Department of Defense, "Department of Defense Directive 3000.09: Autonomy in Weapon Systems, November 21, 2012," Homel. Secur. Digit. Libr., 2012.

[2] S. Davies, "Just War," Eng. Technol., vol. 6, no. 8, pp. 38–40, Sep. 2011, doi: 10.1049/et.2011.0802.

[3] S. Vallor, "The future of military virtue: Autonomous systems and the moral deskilling of the military," Int. Conf. Cyber Conflict, CYCON, 2013.

[4] A. Guersenzvaig, "Autonomous Weapon Systems: Failing the Principle of Discrimination," IEEE Technol. Soc. Mag., vol. 37, no. 1, pp. 55–61, Mar. 2018, doi: 10.1109/MTS.2018.2795119.

[5] P. M. Asaro, "Modeling the moral user," IEEE Technol. Soc. Mag., vol. 28, no. 1, pp. 20–24, 2009, doi: 10.1109/MTS.2009.931863.

[6] "Terminator (1984)," IMDB. https://www.imdb.com/title/tt0088247 (accessed Mar. 10, 2021).

[7] S. Davies, "It's war - but not as we know it [autonomous military robotics]," Eng. Technol., vol. 4, no. 9, pp. 40–43, May 2009, doi: 10.1049/et.2009.0907.

[8] R. C. Arkin, "Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture part I: Motivation and philosophy," in Proceedings of the 3rd international conference on Human robot interaction - HRI '08, 2008, vol. 171, no. 1, p. 121, doi: 10.1145/1349822.1349839.

[9] N. Sharkey, "Cassandra or False Prophet of Doom: AI Robots and War," IEEE Intell. Syst., vol. 23, no. 4, pp. 14–17, Jul. 2008, doi: 10.1109/MIS.2008.60.

[10] D. Callahan, P. Shakarian, J. Nielsen, and A. N. Johnson, "Shaping operations to attack robust terror networks," Proc. 2012 ASE Int. Conf. Soc. Informatics, Soc. 2012, pp. 13–18, 2012, doi: 10.1109/SocialInformatics.2012.22.

[11] J. Khurshid and Hong Bing-rong, "Military robots - a glimpse from today and tomorrow," in ICARCV 2004 8th Control, Automation, Robotics and Vision Conference, 2004., 2004, vol. 1, no. December, pp. 771–777, doi: 10.1109/ICARCV.2004.1468925.

[12] "Just War Theory," Internet Encyclopedia of Philosophy. https://iep.utm.edu/justwar (accessed Mar. 10, 2021).

[13] P. M. Asaro, "How just could a robot war be?," Front. Artif. Intell. Appl., vol. 175, no. 1, pp. 50–64, 2008.

[14] P. W. Kahn, "The Paradox of Riskless Warfare," Philos. Public Policy Q., vol. 22, no. 3, pp. 1–24, 2002.

[15] R. Edson, "Systems thinking. Applied. A primer," Asyst Inst., 2008.

[16] R. H. Spector, "The Vietnam War and the media," Britannica. https://www.britannica.com/topic/The-Vietnam-War-and-the-media-2051426 (accessed Mar. 10, 2021).

[17] "Collateral Murder, 5 Apr 2010," WikiLeaks. https://wikileaks.org/wiki/Collateral_Murder,_5_Apr_2010 (accessed Mar. 10, 2021).

 ISBN: 978-1-61208-838-9