Swedish Defence University

When we see something that is well beyond our understanding

The duty of States to investigate war crimes and how it applies to autonomous weapons systems

Author: Conrad Palmcrantz

Bachelor thesis in the Law of Military Operations

Swedish Defence University

Examiner: Prof. Dr. Jann K. Kleffner

Supervisor: Prof. Dr. Heather Harrison Dinniss

Due date: 9th of January 2019

Word count (excluding bibliography and footnotes): 9073

Table of Contents

List of Abbreviations

AI	Artificial intelligence
API	Protocol Additional to the Geneva Conventions of 12 August 1949 (n 11)
CCW	Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons (n 77)
ECtHR	The European Court of Human Rights
GC's	The Geneva Conventions of 12 August 1949 (n 11)
ICC	The International Criminal Court
ICJ	The International Court of Justice
ICTY	The International Tribunal for the Prosecution of Persons Responsible for Serious Violations of International Humanitarian Law Committed in the Territory of the Former Yugoslavia since 1991
IHL	International humanitarian law
IHRL	International human rights law
ICRC	International Committee of the Red Cross
ILC	International Law Commission
LAWs	Lethal autonomous weapons systems
UNGA	United Nations General Assembly

1. Introduction

1.1 When we see something well beyond our understanding

Russian chess master Garry Kasparov versus IBM supercomputer Deep Blue is a classic match. When Deep Blue won the 1997 rematch, it was a pivotal moment in the development of artificial intelligence (AI). A machine had outmanoeuvred a human and Kasparov himself admitted that this fact made him very afraid: "I am a human being, you know. . . When I see something that is well beyond my understanding, I'm scared." [1]

Fast-forward to 2016 and a similar event of man versus machine takes place. This time, the board game of choice was the ancient Chinese pastime called Go.[2] Google DeepMind put their algorithm AlphaGo to the test by challenging world champion Lee Sedol to a five-game match. The size of the board and the number of possible moves make Go far more complicated than chess, and it is virtually impossible for a computer to conduct an exhaustive search of every conceivable game plan.[3] Go requires a great deal of creativity and intuition, not only the brute-force calculations Deep Blue relied upon to beat Kasparov, and therefore machine learning is indispensable.[4]

Because of the complexity of Go, Sedol was fairly certain he would win, and the DeepMind team was intensely worried they would lose and look foolish in the process.[5] However, to much surprise, AlphaGo won 4-1 and even more surprising was *how* it had won. In one of the games, AlphaGo played a series of lazy-looking defensive moves that prompted commentators to suppose that it had malfunctioned and the humans in the room all agreed that it was a "weird" tactic.[6] Later, after AlphaGo had won, the DeepMind team confessed that they were not good enough Go-players to explain the defensive moves and could not say with certainty why AlphaGo behaved like that.[7]

This example shows that it is difficult to unpick the rationale of an algorithm. The anecdote may seem insignificant as it only concerns a board game, but it has been argued

---

[1] Charles Krauthammer, 'Be Afraid' [1997] The Weekly Standard <https://www.weeklystandard.com/charles-krauthammer/be-afraid-9802> accessed 1 November 2018.

[2] For a general overview of the event see Greg Kohs, *AlphaGo [Documentary]* (2017).

[3] Kohs (n 2) at 12 mins.

[4] 'AlphaGo' (*DeepMind*) <https://deepmind.com/research/alphago/> accessed 6 December 2018.

[5] Kohs (n 2) at 27 mins.

[6] Kohs (n 2) at 76 mins.

[7] Kohs (n 2) at 78 mins.

that the ideas driving the AlphaGo are the ideas that will drive our entire future.[8] The machine learning logic of the AlphaGo could theoretically be implemented in, for example, self-driving cars, healthcare technology, and e-commerce solutions. However, the potential is not restricted to benevolent technologies, and it may be utilized in a much more controversial field of technology: Lethal autonomous weapons systems (LAWs).

Although a fully autonomous weapons system operating in a complex environment without human supervision is not currently within reach, efforts are being made to enhance autonomous capabilities in weapon systems.[9] For example, the United States funds research on automatic target recognition from aerial platforms utilizing machine learning.[10] This is cause for concern. If humans have trouble understanding the intricacies of AlphaGo, a narrow AI system applied to a board game, it certainly seems impossible to explain a broad AI system acting on the battlefield. If a complex war-algorithm acts "weird" and engages in destructive behaviour, the question arises how humans should react. When we see something beyond our understanding, what ought we do?

1.2 Research aim

1.2.1 Research question

The purpose of this thesis is to examine how the duty of States to investigate potential war crimes applies to incidents involving LAWs. War crimes will be narrowly understood as grave breaches of the Geneva Conventions (GC's) and its Additional Protocol I (API).[11]

[8] Metz, Cade in Kohs (n 2) at 13 min.

[9] For a general overview of what technology is presently available see e.g. Vincent Boulanin and Maaike Verbruggen, *Mapping the Development of Autonomy in Weapon Systems*, vol 2017.

[10] 'Automatic Target Recognition of Personnel and Vehicles from an Unmanned Aerial System Using Learning Algorithms | SBIR.Gov' <https://www.sbir.gov/sbirsearch/detail/1413823> accessed 4 October 2018.

[11] Geneva Convention for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field, 12 August 1949, 75 UNTS 31 (GC-I), art 49; Geneva Convention for the Amelioration of the Condition of Wounded, Sick and Shipwrecked Members of Armed Forces at Sea of August 12, 1949, 75 UNTS 85 (GC-II) art 50; Geneva Convention Relative to the Treatment of Prisoners of War of August 12, 1949, 75 UNTS 135 (GC-III) art 129; Geneva Convention Relative to the Protection of Civilian Persons in Times of War of August 12, 1949, 75 UNTS 287 (GC-IV) art 146; Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977, 1125 UNTS 3 (AP-I), Article 85.

Regarding autonomous technology, I will focus specifically on *deep reinforcement learning*.[12] In order to fulfil the research aim, the following questions will be studied:

- What are the main difficulties associated with interpreting technology that employs deep reinforcement learning?
- What incidents of alleged breaches must be investigated?
- How is the duty to investigate triggered?
- What are the relevant investigative standards?
- Is the current legal framework efficient when applied to LAWs employing deep reinforcement learning?

1.2.2 Limitations

There are indeed other core crimes of international law that States must investigate besides the grave breaches found in the GC's and API.[13] However, the aim of this thesis is *not* to meticulously discuss States' duty to investigate transgressions of international humanitarian law (IHL). Instead, only a few examples of war crimes will be presented and set the stage for an in-depth discussion on autonomous weapons, investigative standards and accountability for breaches of IHL. A further inquiry into other violations would surely have been interesting, but falls outside the scope of this thesis.

Moreover, the reason for limiting the technological inquiry to the concept of deep reinforcement learning is because it is generally considered to be a powerful application of artificial intelligence. A member of the Google Deepmind team went as far as suggesting this formula: artificial intelligence = reinforcement learning + deep learning.[14] I am open to the idea that LAWs may eventually utilize another type of logic, but the current state of technology indicates that deep reinforcement learning is the most feasible approach.

[12] This concept is explained in section 2.

[13] Jann K Kleffner, *National Suppression of Core Crimes* (Oxford University Press 2008) ch 2.1.

[14] David Silver, 'Deep Reinforcement Learning' <http://videolectures.net/rldm2015_silver_reinforcement_learning/> accessed 30 November 2018, at 2 min.

1.3 Methodology

1.3.1 How I will approach the law

A legal doctrinal method will be employed to answer the research questions. As previously underlined, the GC's and API are the primary legal authorities of interest. Nevertheless, to make sense of these documents, it is essential to consult other sources of international law. Manifestly, how State parties interpret and operationalize their obligation to investigate grave breaches is heavily influenced by practices of international organisations and international tribunals.[15] Regarding investigative standards, international human rights law (IHRL) plays a significant part, and this prompts a brief discussion on the interaction between IHRL and IHL

In the *Nuclear Weapons* advisory opinion[16] and the *Israeli Wall* advisory opinion,[17] the International Court of Justice (ICJ) ruled that both IHRL and IHL applies in times of war. The impact of certain IHRL norms varies depending on the subject matter, and the ICJ's opinions suggest that IHRL should be treated as *lex generalis* regarding the conduct of hostilities, whereas the more particular IHL framework should be dealt with as *lex specialis*. Furthermore, relating to the prosecution of war crimes, the International Criminal Tribunal for the former Yugoslavia (ICTY) has made circumstantial interpretations of IHRL requirements in the context of armed conflict.[18]

Specifically, regarding State's obligation to investigate, Professor Michael Schmitt has argued that investigative standards are circumstantial and that a State's ability to investigate crimes is severely impaired in times of war.[19] For instance, evidence may have been destroyed in battle, travel is dangerous, judicial bodies are usually located far away from

[15] For a general discussion on the internationalization of domestic procedures, see e.g. Goran Sluiter, 'Law of International Criminal Procedure and Domestic War Crimes Trials' [2006] International Criminal Law Review 605.

[16] *Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion*, ICJ Rep 1996, p. 226, para 25.

[17] *Legal Consequences of the Construction of a Wall in the Occupied Palestinian Territory, Advisory Opinion*, ICJ Rep 2004, p. 136, paras 111-112.

[18] *Tadic case* (Decision on the Prosecutor's motion requesting protective measures for victims and witnesses) ICTY-94-1 (10 August 1995), paras 17-30; *Celebici case* (Decision on the motions by the Prosecution for protective measures for the prosecution witnesses pseudonymed 'B' through to 'M') ICTY-96-21 (28 April 1997), para 27.

[19] Michael N Schmitt, 'Investigating Violations of International Law in Armed Conflict' in *Essays on Law and War at the Fault Lines* (TMC Asser Press 2011), 607.

the battlefield, and standard forensic tools may be unavailable. IHL norms are developed with those predicaments in mind, and this strongly indicates that IHL norms are *lex specialis*.

Nevertheless, a thorough assessment of specific norms is still necessary, and it is impossible to categorically exclude that an IHRL norm could be more specific in certain situations. Consequently, IHL norms on the investigation of war crimes will principally be treated as a special application of the general IHRL requirements. IHRL will largely serve a complementary function by providing interpretive guides to the more special rules and by filling gaps in the IHL regulations.[20] Efforts will be made to harmonize disputing norms, but if an apparent conflict of norms arises, IHL will – most likely, but not necessarily always – be given precedence by virtue of *lex specialis*.[21]

1.3.2 How I will approach the technology

There is no treaty definition of what is and what is not a LAWs. Countless descriptions have been suggested, each with its pros and cons. A frequently referenced definition is one by the United States' Department of Defense:

> [LAWs is] a weapon system that, once activated, can select and engage targets without further intervention by a human operator. This includes human-supervised autonomous weapon systems that are designed to allow human operators to override operation of the weapon system but can select and engage targets without further human input after activation.[22]

What this definition fails to acknowledge is that autonomy could be utilized in support of capabilities other than targeting. Examples thereof are autonomy in mobility,[23] interoperability,[24] and intelligence gathering.[25] Since the nature of autonomy may vary depending on the specific capability, the degree of human control and the level of program sophistication, a 'functional approach' has been advised.[26] This approach focuses on autonomy as a spectrum in relation to specific tasks and not as a fixed general concept.[27]

---

[20] See e.g. ILC rep, 'Fragmentation of international law: Difficulties arising from the diversification and expansion of international law' (1 May-9 June and 3 July-11 August 2006) A/CN.4/L.682, paras 98-102.

[21] *Ibid.* paras 103-107.

[22] U.S. Department of Defense, 'Directive 3000.09, Autonomy in Weapon Systems'

[23] Boulanin and Verbruggen (n 9) 21.

[24] Boulanin and Verbruggen (n 9) 29.

[25] Boulanin and Verbruggen (n 9) 27.

[26] Boulanin and Verbruggen (n 9) 7.

[27] Boulanin and Verbruggen (n 9) 6–7.

As previously mentioned, a fully autonomous system operating in a complex environment without human supervision is not currently within reach.[28] Furthermore, States may demand, as a matter of law or policy, that a human operator always remains in the decision loop and the idea of upholding *meaningful human control* is common in the discussion on LAWs.[29] Because of this interplay between technology and policy, it is impossible to say with certainty what lies ahead. One can imagine a situation where a supervising operator interacts with a deep reinforcement learning LAWs, as well as a fully autonomous weapon acting independently. This uncertainty causes methodological concerns since a functional approach is best suited for technology that is already in use. However, a functional approach does not *per se* exclude a discussion on what technology may be available in the future.

When venturing into the unknown, it is a significant risk that one applies the law on sci-fi narratives and to avoid this fallacy, I will apply a method suggested by CopeTech, a joint effort Swedish Royal Institute of Technology and Swedish Defence Research Agency (FOI).[30] According to this methodological framework, one should consider different potential scenarios and acknowledge that society will react to the technological development, which will alter the technological trajectory in a co-evolutionary process. [31] Furthermore, when developing the co-evolutionary scenarios, it is essential to have a specific technological object in mind.[32]

Thus, what I set out to do is to focus on a specific technology – namely, deep reinforcement learning – and consider the different ways it may be utilized in future weapon capabilities. This assessment will then serve as the object upon which I will apply the existing legal framework and discuss potential issues – a legal reaction in the co-evolutionary scenario, so to speak. Sticking to the very basics of machine learning and big data technology, I hope to anchor my thoughts in feasible predictions.

---

[28] For a general overview see e.g. Boulanin and Verbruggen (n 9).

[29] See e.g. Thompson Chengeta, 'Defining the Emerging Notion of Meaningful Human Control in Weapon Systems' (2016) 49 New York University Journal of International Law and Politics 833.

[30] Henrik Carlsen and others, 'Assessing Socially Disruptive Technological Change' (2010) 32 Technology in Society 209.

[31] Carlsen and others (n 30).

[32] For a summary in Swedish see Linda Johansson, *Äkta robotar* (Fri Tanke Förlag 2015) 162.

1.4 Thesis outline

Section two will frame the problem by explaining the technological context of deep reinforcement learning. *Section three* will introduce the grave breaches regime, examine how the duty to investigate is triggered, and discuss potential problems relating to deep reinforcement learning LAWs. *Section four* will analyse the duty to investigate under command responsibility and apply it in the technological context. *Section five* will examine how investigations into grave breaches should be conducted and what bearing those principles have on military operations involving deep reinforcement learning LAWs. *Section six* will provide a few closing remarks.

2. Framing the problem: Deep reinforcement learning as a black-box

2.1 Introduction

AlphaGo was a breakthrough for a specific type of machine learning called *deep reinforcement learning.*[33] When setting up the AlphaGo, the human programmers did not explicitly tell it how to play, because that would require extreme computing power in processing the information and extreme man-power in coding the system.[34] Instead, the AlphaGo mimics human intuition through deep reinforcement learning, a technique integrating *reinforcement learning* with *neural networks.*[35] In the following section, I will describe how this technology creates a black-box that is difficult to investigate.

2.2 Machine learning and deep neural networks

Machine learning is an algorithm that can extract patterns from data.[36] If a computer program can improve its performance of a task with experience, it is employing machine learning.[37] For this to be possible, a task needs to be defined, for example recognizing

[33] Yuxi Li, 'Deep Reinforcement Learning: An Overview' (2017) <http://arxiv.org/abs/1701.07274> accessed 30 November 2018.

[34] Kohs (n 2) at 12 mins.

[35] *Ibid.*

[36] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning* (The MIT Press 2016) 2.

[37] *Ibid.* 96–97.

imagery of a particular object.[38] Then, quantitative measures of its performance must be designed, for example the error rate when classifying images.[39] Lastly, the algorithm must gain experience by being fed data, such as imagery. [40]

Deep neural networks are one example of possible machine learning software architecture. They are referred to as *neural* because they are loosely inspired by how the human brain functions.[41] The network consists of units that cooperate much like neurons do in the brain: The neuron-unit receives input signals and calculates (weights) the information, leading to a new representation of the information as an output signal.[42]

These systems are labelled *networks* because they consist of different interconnected functions, feeding information forward from input to output.[43] The overall goal of the network is to approximate a specific output function, and this function is constructed as a chain of other more particular functions.[44] Each link of the chain is referred to as a network *layer*, representing a function and consisting of neuron-like units.[45] The overall length of the chain, i.e., the number of layers stacked together, determines the depth of the neural network and a *deep* neural network contains *hidden layers*. [46]

The final layer has a result we can observe.[47] The first layer contains input (the data set) that also is detectable for humans.[48] The behaviour of the other layers in-between, however, are not observable and how the neuron-units weight the input information is not specified in the training data.[49] By modifying the connections between the neuron-units, i.e., how the inputs are weighted, the algorithm can enhance its performance.[50] These modifications are determined by the algorithm itself, and the output of the middle layers is unknown – or hidden, as the name suggests.[51]

---

[38] *Ibid.* 97–99.
[39] *Ibid.* 100–101.
[40] *Ibid.*
[41] *Ibid.* 164.
[42] *Ibid.*
[43] *Ibid.* 163.
[44] *Ibid.*
[45] *Ibid.*
[46] *Ibid.* 164.
[47] *Ibid.*
[48] *Ibid.* 6.
[49] *Ibid.* 164.
[50] Andreas Matthias, 'The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata' (2004) Ethics and Information Technology 175, 178.
[51] Goodfellow, Bengio and Courville (n 36) 164.

2.3 Deep reinforcement learning

Reinforcement learning is a process of trial-and-error, in which the system learns from rewards and punishments.[52] The system interacts with its environment and learns from experience it gains after being deployed in its final operating setting.[53] In this kind of learning, there is no clear distinction between the training of the algorithm and the application of the algorithm.[54] An algorithm instructed to maximize the reward in a game, for example, could estimate the reward of a move based on samples from moves it has made previously during its deployment.[55] Accordingly, the system could learn from an offline memory bank it has required after deployment through replaying experience.[56]

More specifically, the learning process occurs in state-outcome pairs.[57] The physical presence of the system is called an agent, for example a drone out on a mission.[58] The state is the surroundings of an agent and the action is whatever the agent decides to do.[59] After an action, the agent observes how the environment reacts to its action and then calculates the reward of the action (the value function).[60] In *deep* reinforcement learning, calculating the value function is done by applying a deep neural network.[61] Simultaneously as the agent calculates the reward, it observes the environment to initiate a new state-outcome pair.[62]

2.4 Why it is a Black-box

A suggested definition of the term black-box is: 'A usually complicated electronic device whose internal mechanism is usually hidden from or mysterious to the user.'[63] That is a fitting description of an autonomous system that relies on deep reinforcement learning. The structure of the system is inherently complicated and offers few explanations to its

---

[52] 'Deep Reinforcement Learning' (*DeepMind*) <https://deepmind.com/blog/deep-reinforcement-learning/> accessed 30 November 2018.

[53] Matthias (n 50) 179.

[54] *Ibid.* 179.

[55] Volodymyr Mnih and others, 'Playing Atari with Deep Reinforcement Learning' <https://arxiv.org/abs/1312.5602> accessed 29 November 2018.

[56] Li (n 33) 16.

[57] *Ibid.* 31.

[58] 'A Beginner's Guide to Deep Reinforcement Learning' (*Skymind*) <http://skymind.ai/wiki/deep-reinforcement-learning> accessed 29 November 2018.

[59] 'A Beginner's Guide to Deep Reinforcement Learning' (n 58).

[60] Li (n 33) 9.

[61] 'Deep Reinforcement Learning' (n 52).

[62] 'A Beginner's Guide to Deep Reinforcement Learning' (n 58).

[63] 'Definition of BLACK BOX' <https://www.merriam-webster.com/dictionary/black+box> accessed 29 November 2018.

decisions. The hidden parameters are mysterious, even for computer scientists, who admit the lack of interpretability limits the development of deep reinforcement learning.[64]

To put it simply, we only see the input and the output.[65] Tracing every single activation of each neuron-unit throughout the learning process would create an audit trail that is incomprehensible for a human being.[66] Additionally, the reward signal may be varied, delayed, or affected by unknown variables in the environment, which taint the feedback loop in state-outcome pairs.[67] Consequently, human programmers lack the necessary tools to fully analyse *what* an agent has learned and *why* it has learned it.[68]

2.5 Why it is legally relevant

Despite the apparent technological intricacies, there is a tendency to simplify the issue of deep learning. At the 2014 informal meeting of expert on LAWs, organized within the framework of the Convention on Certain Conventional Weapons (CCW),[69] the United States delegation declared that:

> There remains a lack of clarity regarding the concept of autonomous weapons decision making. As we have said, it is important to remind ourselves that machines do not make decisions; rather, they receive inputs and match those against human programmed parameters.[70]

That is a clear example of a generalization that would not be entirely true for a LAWs utilizing deep reinforcement learning.[71] The parameters in such a system – the calculation of input data in hidden layers – would not be directly programmed by a human operator and could change after interactions in the operative environment through reinforcement

[64] Li (n 33) 5.

[65] 'A Beginner's Guide to Deep Reinforcement Learning' (n 58).

[66] Ariel Bleicher, 'Demystifying the Black Box That Is AI' (*Scientific American*) <https://www.scientificamerican.com/article/demystifying-the-black-box-that-is-ai/> accessed 27 November 2018.

[67] 'A Beginner's Guide to Deep Reinforcement Learning' (n 58).

[68] Tom Zahavy, Nir Ben Zrihem, and Shie Mannor, 'Graying the Black Box: Understanding DQNs'

[69] Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May be Deemed to be Excessively Injurious or to Have Indiscriminate Effects (As Amended on 21 December 2001), 10 October 1980, 1342 UNTS 137

[70] Statement by the United States at the 2014 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS), May 16, 2014 (as cited in Chengeta (n 29) 859)

[71] To be fair, there has been a paradigm shift in AI research recently, and it is plausible that the US representative had another AI approach in mind. See e.g: Anna Lee Strachan, *Can We Build a Brain? [Documentary]* (2018) at 10 min.

learning. One of the great benefits of a deep learning algorithm is that programmers abstain from deciding on relevant parameters, allowing the algorithm to identify the most efficient solution.[72] Of course, there are varying degrees of human supervision in selecting the training data and choosing regularization strategies for optimizing performance.[73] Still, the raison d'être of deep learning is that it chooses its preferences and makes manual coding unneeded in significant functions.[74]

Returning to the example of AlphaGo, it is noteworthy that no one told it how to play or what tactics to use in order to beat Lee Sedol. AlphaGo had to figure it out autonomously and the end result surprised its human programmers. Imaginably, a LAWs could act equally surprising, and since deep reinforcement learning functions as a black-box, it may be impossible to review what has happened precisely. That is problematic considering State's duty to investigate violations of IHL.

3. The State's duty to investigate

3.1 Introduction

After describing the technological basics and framing the problem, I will now direct my attention to the relevant legal provisions regarding the duty to investigate suspected incidents of war crimes under the *grave breaches regime*. In this section, I will briefly explain the necessary elements of a grave breach and how the duty to investigate is triggered. This will subsequently be discussed in relation to the complexities of deep reinforcement learning LAWs.

3.2 Presenting the Grave Breaches Regime

After the atrocities of World War II, the international community established a binding obligation to penalize serious wrongdoings in wartime. In 1949, State parties agreed on the four Geneva conventions, which are applicable in situations of international armed conflict between parties to the convention.[75] In each convention, there is an identical article on the investigation of grave breaches:

[72] Bleicher (n 66).
[73] Goodfellow, Bengio and Courville (n 36) 221.
[74] Boulanin and Verbruggen (n 9) 17.
[75] GC's common article 2. NB. Although this thesis focuses on international armed conflicts, certain grave breaches are also prohibited in non-international armed conflicts.

> Each High Contracting Party shall be under the obligation to search for persons alleged to have committed, or to have ordered to be committed, such grave breaches, and shall bring such persons, regardless of their nationality, before its own courts. It may also, if it prefers, and in accordance with the provisions of its own legislation, hand such persons over for trial to another High Contracting Party concerned, provided such High Contracting Party has made out a prima facie case.[76]

This provision obligates State parties to pursue allegations of grave breaches. They must either prosecute or extradite persons accused of grave breaches (*aut dedere, aut judicare).*[77] Subsequent treaty articles define the central notion of grave breaches, and the GC's contain common provisions on the constitutive elements.[78] These have later been reiterated and developed in API,[79] and examples of grave breaches include:

- Wilful killing[80]
- Torture or inhumane treatment, including biological experiment[81]
- Wanton destruction[82]
- Indiscriminate attack affecting civilians[83]
- Making civilians the object of attack[84]
- Perfidious use of protective signs[85]

The various breaches have different material elements *(actus reus*) and mental element (*mens rea).* However, concerning all offences, there are two common constitutive material elements. Firstly, there must be a link (a *nexus)* between the act and the state of belligerency.[86] Secondly, the act must be committed against a protected person or protected property.

Parts of the relevant case law on grave breaches has been developed in response to non-international armed conflicts and will be duly considered.

[76] GCIV art 146 (n 7).

[77] Schmitt, 'Investigating Violations of International Law in Armed Conflict' (n 19) at 592.

[78] GCI, art 50, GCII, art 51; GCIII, art130; GCIV, art 147. (n 7)

[79] API, art 85-91 (n 7).

[80] GCI art 50 (n 7)

[81] GCI art 50 (n 7)

[82] GCI art 50 (n 7)

[83] API, art 85.3 (n 7)

[84] API, art 85.3 (n 7)

[85] API, art 85.3 (n 7)

[86] ICRC, 'Commentary on the First Geneva Convention' (Cambridge University Press 2016) para 2922.

3.3 The two common constitutive material elements

The definition of protected persons and property causes no obvious investigative issues that are specific to LAWs. The protected persons remain the same as defined in the GC's and API.[87] Regarding protected property, no definition is provided as such. Rather the GC's and API define what objects cannot be attacked, and those objects are protected.[88] That definition equally applies to LAWs and investigating against whom or what an attack was directed can be done using a standard procedure. Destruction resulting from the deliverance of kinetic force is detectable for human investigators, regardless if the weapon is autonomous or conventional.

Regarding the *nexus* requirement, however, there are specific problems relating to LAWs. Concerning grave breaches, the ICTY has ruled in the *Kunarac* case that the armed conflict, at a minimum, must have played a substantial part in the perpetrator's ability to commit the act, deciding on the act, the manner of conducting the act, or the purpose of the act.[89] Relevant indicative factors could be the perpetrator's status as a combatant, the victim's status as a combatant, the ultimate goal of the military campaign, and the official duties of the perpetrator.[90] The indicative factors can be applied to members of the armed forces and civilians alike since potential perpetrators are not limited to any specific group.[91]

In the context of LAWs, the *nexus* requirement may cause various degrees of complications, depending on the level of human involvement. If there is a human operator who exercises substantial control, the situation is less complicated since that person is just as investigable as a person operating a conventional weapon. The *Kunarac*-criteria could be applied to the operator and the challenges for the investigator, in such case, are comparable to the difficulties faced when investigating incidents involving remotely piloted weapon platforms.

A more problematic subject is fully autonomous weapons, where there would not be any individual directing the use of force. If we imagine a LAWs which is so fast that no human can intervene, or which is capable of operating when the lines of communication have broken down, it is hard to see how a human operator would be able to exert

[87] See e.g. GCI art 13; GC II art 4; GCIV art 4; API arts 8, 11, and 85 (n 7).
[88] ICRC (n 86) para 2928.
[89] *Kunarac* case (Appeals judgement) ICTY IT-96-23& IT-96-23/1-A (12 June 2002) para 58
[90] *Kunarac* case (n 89), para 59
[91] ICRC (n 86) para 2929.

significant influence. In such a scenario, weapon developers would play an increasingly important role and correspondingly become an increasingly import actor to investigate.

Indeed, the work of weapon developers has already been scrutinized by the authorities during the legal review of new weapons. It follows from art 36 of API that new means of warfare, which undoubtedly includes new LAWs, must be tested and the nature of the weapon must be deemed in compliance with IHL before deployment in battle. However, when investigating an incident *ex post facto*, the *nexus* requirement may exclude the application of the grave breaches regime and consequently not give rise to a separate duty to investigate weapons developers.

Before applying the *Kunarac*-criteria to the weapon developers, it is important to note that it is a diverse group of people. LAWs are often developed through a joint effort between State agencies and private companies.[92] The involved actors could be civilian computer scientists without any military aspiration as well as service members of the armed forces instructing the programmers. To further complicate matters, the development would presumably take place far away from the theatre of war and start before the commencement of hostilities. However, it is conceivable that development could continue during an armed conflict since programmers might tweak the algorithm to enhance performance.

If the development of the LAWs took place before the conflict broke out, it is probable that the grave breaches regime does not apply.[93] Regarding actions taken in peacetime without any particular armed conflict in mind, there is a lack of case law to support that the *nexus* requirement could be met. Commercial actors have been convicted of war crimes, for example German businessmen who plundered occupied territories or who provided the Third Reich with Zyklon B gas,[94] but in those cases the culpable actions were taken during an already ongoing armed conflict. The development of LAWs in peacetime lacks that contextual element, and according to the general prohibition of analogies in criminal

[92] For example, Google develops and supports algorithms that the United States' Department of Defense requires to analyse imagery from military drones 'Google Ditches Department of Defense, Updates Its Code of Ethics' (*Futurism*) <https://futurism.com/maven-google-military-tech> accessed 30 November 2018.

[93] This conclusion is supported by Tim McFarland and Tim McCormack, 'Mind the Gap: Can Developers of Autonomous Weapons Systems Be Liable for War Crimes' (2014) 90 International Law Studies Series. 378.

[94] *Röchling case,* Superior Military Government Court of the French occupation zone in Germany, 30 June 1948, *Trials of War Criminals before the Nuremberg Military Tribunals Under Control Council Law No. 10,* vol 14 p. 1119; *Zyklon B case,* British Military Court, case no 9, 1-8 March 1946, Hamburg, *I Law reports of Trial of War criminals 93, 94* p. 103.

law (*nullum crimen sine lege*), the elements of the crime must be interpreted narrowly. Consequently, these acts would fall outside the scope of the grave breaches regime and do not entail a separate duty to investigate.

Nevertheless, if the development continues during an armed conflict, for example by maintaining the system and updating its software, a sufficient link could be established. The weapons developer could reasonably foresee that the actions of the LAWs would be conducted against enemy combatants and enemy military objects. The developer would have a position within the military system, perhaps not as a combatant but maybe as some sort of technical advisor to the military. Rationally, developers would be informed about the ultimate goal of the military campaign and adjust the capability of the LAWs thereafter. Furthermore, the coding would be an official duty regulated in an employment agreement or service contract between the State and a private company. It is therefore feasible that the state of belligerency could play a substantial part in the weapon developers' ability to program – intentionally or by mistake – a LAWs that later is involved in a grave breach of IHL. This indicates that the duty to investigate grave breaches could include inquiries into how a LAWs is programmed during an armed conflict.

3.4 The mental element

Finally, each material element must be covered by a sufficient mental element. It is a general principle of law that an act is not culpable unless the mind is guilty (*actus reus non facit reum nisi mens sit rea*). However, there is no uniform rule on the mental element applicable to all grave breaches and courts tend to establish the required level of *mens rea* on a case-by-case basis.[95] For some breaches, the treaty provision indicates the required level: A murder must be *wilful*, whereas destruction of property is a grave breach if it is done *wantonly*.[96] Furthermore, the mental element may vary depending on the mode of liability.[97]

In lack of specification on the mental element, one can look to art 30 of the Rome statute for guidance.[98] This article, arguably reflective of customary law,[99] generally applies

[95] Robert Cryer and others, *An Introduction to International Criminal Law and Procedure* (Cambridge University Press 2010) ch 15.7.

[96] GCI art 50 (n 7)

[97] Cryer and others (n 95) 385.

[98] Rome Statute of the International Criminal Court (17 July 1998) UN Doc A/CONF.183/9.

[99] Cryer and others (n 95) 385.

to all international crimes, including grave breaches, and requires that the actions are committed with intent and knowledge. What this entails is, firstly, that the person must mean to engage in the conduct of the crime. Secondly, if the material element of the crime requires a particular consequence, the person must intend to cause the consequence or as a minimum know that the consequence will occur in the ordinary course of events.[100]

Applying this principle to deep reinforcement learning LAWs is a massive challenge. If, for example, the operators blame an action on the agent's ability to autonomously adapt its behaviour to the operative environment, it is difficult for an investigator to evaluate the credibility of that defence. Grasping the logic of deep reinforcement learning is not always technologically possible, which may prove to be an insurmountable obstacle. To some extent, circumstantial evidence could establish the mental element. It is possible to gather *indicia* and construe a realistic account of events.[101] Colleagues could testify that an individual operator or weapons developer consciously disregarded the risk of violating IHL, internal reports may be leaked, and there could be apparent biases in the training data, for example. That would at least enable an initial investigation of the mental element.

Still, the black box issue remains a constant source of doubt. Deep reinforcement learning LAWs behave in such a way that it is problematic to establish what will occur in the ordinary course of events. Since no human can explain the exact logic of each activation of the neuron-units in hidden layers, the LAWs might act unforeseeably and cause extraordinary consequences. Additionally, given the logic of reinforcement learning, the LAWs would continuously alter its behaviour depending on what is rewarded in battle, and human programmers may not be able to assess this conduct fully. The ICC has ruled that art 30 is a standard of 'virtual certainty' or 'practical certainty,'[102] but the advanced technology makes it integrally arduous to investigate what is certain. The human actors involved in an incident could blame the complex algorithms and claim that they had no criminal intent. That is a novel challenge for States when investigating potential breaches.

[100] Art 30(2), Rome Statute (n 98)

[101] The admissibility of such evidence depends on domestic procedural law. In international fora, it is commonly used. See e.g. *Hadzihasanovic and others case* (Trial judgement) ICTY-01-47 (15 March 2006) para 94.

[102] *Bemba case*, (Confirmation of charges) ICC-01/05-01/08, (15 June 2009), para 362

3.5 Triggering the duty to investigate

There is no duty in the GC's and API triggering a proactive duty to uncover IHL violations.[103] Although States are obligated to enact effective national legislation to suppress war crimes, there is no explicit IHL requirement amounting to a pre-emptive duty to search for grave breaches.[104] Instead, the triggering factor is that an incident has come to a State party's attention.[105] Examples of relevant information are accusations from victims, NGO:s reporting violations, a request to extradite a suspect, or stories in the news media regarding suspected grave breaches.

How reliable the information must be to trigger an investigative duty is not unambiguously regulated. When a State is requested to extradite a suspect, the petitioning State must, according to the GC's, make out a *prima facie* case.[106] National legislation limits when extradition is allowed and the interpretation of what is a *prima facie* case varies, but as a general rule, it would require evidence that typically would lead to penal prosecution in domestic courts.[107] Regarding other allegations, the GC's and API are mute, but it has been argued that allegations must give rise to a *credible* suspicion that a breach of IHL has occurred.[108] The approach of the United States could be said to embody this standard, and according to American domestic regulations, a preliminary review of the 'totality of the circumstances' must give rise to a particularized basis for suspecting a violation of the laws of war.[109]

In contrast, Swedish law on disciplinary misconduct within the military,[110] sets the bar slightly lower. The supplementary governmental regulations stipulate that responsible

---

[103] Orna Ben-Naftali and Roy Peled, 'How Much Secrecy Does Warfare Need?' in Andrea Bianchi and Anne Peters (eds), *Transparency in International Law* (Cambridge University Press 2013) 354.

[104] Ben-Naftali and Peled (n 103) 354.

[105] See e.g. *Isayeva, Yusupova and Bazayeva v. Russia* (Judgement) ECHR application Nos. 57947/00, 57948/00, 57949/00) (24 February 2005) paras 209–213.

[106] See the common wording of the GC's (n 7).

[107] Jean S Pictet (ed), *Commentary: Fourth Geneva Convention Relative to the Protection of Civilian Persons in Time of War* (ICRC 1958) 593.

[108] See e.g. Schmitt, 'Investigating Violations of International Law in Armed Conflict' (n 19) 628.

[109] Dick Jackson, 'Reporting and Investigation of Possible, Suspected, or Alleged Violations of the Law of War' [2010] Army Lawyer 99. Jackson suggests an analogous application of 'reasonable suspicion' as articulated in *Terry v. Ohio*, U.S. Supreme Court. June 10, 1968. 392 U.S. 1.

[110] [Swedish law re disciplinary misconduct] *Lag (1994:1811) om disciplinansvar inom totalförsvaret, m.m*

military authority must investigate any disciplinary misconduct (*disciplinförseelse)* if a person under responsible command is accused thereof, or if the organisation otherwise receives information about possible misconduct.[111] Violations of internal policy, as well as violations of national and international law, must be investigated.[112] At this point, the legal standard of proof requires that disciplinary misconduct may be presumed (*kan antas*),[113] which is generally considered to be the lowest standard of proof in Swedish law.[114] Weak suspicions of an act that mostly fits the objective legal elements would suffice, and an individual suspect is not necessary at this stage.[115]

What this means for deep reinforcement learning LAWs is unclear. The current legal provisions are vague, and States may take a position based on policy concerns rather than *opinio juris*. Therefore, it is hard to say what triggering factors stem from the grave breaches regime and what stems from excessively strict State policy. One cannot conclusively declare that the United States' policy violates IHL, nor can one conclude that the Swedish approach is an example of a 'best practice.' States have a considerable margin of appreciation regarding triggering factors, but that could potentially change in case LAWs are fielded. Civil society has expressed worry about unaccountable "killer robots,"[116] and such campaigns may force States to investigate incidents at an earlier stage. States could do so because of policy reasons, e.g., to secure popular support for LAWs, or legal reasons, e.g., to comply with a future multilateral treaty.

4. The commander's duty to investigate

4.1 Introduction

After analysing the State's general duty to investigate, a particular State agent will be scutinized: The commander. Some authors have argued that the doctrine of command

[111] [Swedish government regulation re disciplinary responsibility] *Förordning (1995:241) om disciplinansvar inom totalförsvaret m.m.* chpt 2, para 4.

[112] *Ibid.*

[113] *Ibid.*

[114] See e.g. Christian Diesen, *Bevis 7: Bevisprövning i förvaltningsmål* (Norstedts Juridik AB 2003) at 93.

[115] [Swedish preparatory works re criminal investigation] *Prop 1994/95:23 Ett effektivare brottmålsförfarande,* at 76.

[116] See e.g. 'Campaign to Stop Killer Robots' <https://www.stopkillerrobots.org/> accessed 19 December 2018.

responsibility is a separate source for the legal obligation to investigate violations of IHL.[117] Others claim that command responsibility is a necessary enforcement of the State's general obligation to investigate violations under the GC's and API.[118] Regardless of how one conceptualises the issue, it deserves particular attention and especially so when applied to deep reinforcement learning LAWs.

4.2 Presenting command responsibility

Command responsibility applies broadly from the highest level of political and strategic commanders to low-level squad leaders with only a few subordinates.[119] In this multitude of responsible commanders, the duties fluctuate depending on context and, to exemplify, the responsibilities of a battalion leader differ from the responsibilities of a non-commissioned squad leader.[120] However, generally speaking, responsible command is obligated to scrutinize their subordinates and commanders are usually in a position that enables them to establish basic facts of military operations.[121] The triggering factor of the commander's duty to investigate is not explicitly stipulated. Nevertheless, there is a doctrine of command responsibility, which explains when investigative inaction is culpable. Under customary law, there are three cumulative requirements to establish command responsibility: (1) A superior/subordinate relationship, (2) a sufficient mental element, and (3) a failure to take reasonable steps to prevent or punish breaches.[122]

4.3 Superior/subordinate relationship

The first requirement is tailored for a human chain of command and not a human-machine relationship. A person officially appointed as a military commander is a *de jure* commander, and a person effectively acting as a military commander is a *de facto* commander.[123] The commander must exercise 'effective control' in the sense that he or she must have a

[117] See e.g. Jackson (n 109) 95.

[118] Amy ML Tan, 'The Duty to Investigate Alleged Violations of International Humanitarian Law: Outdated Deference to an Intentional Accountability Problem' 49 International Law and Politics 182.

[119] Claude Pilloud and others, *Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949* (Kluwer Academic Publishers 1987) para 3553

[120] Pilloud and others (n 119) para 3554.

[121] Pilloud and others (n 119) para 3560.

[122] Cryer and others (n 95) 389.

[123] Antonio Cassese, *Cassese's International Criminal Law* (3 edn. Oxford University Press 2013) 188.

material ability to prevent or punish subordinates who violate the law.[124] These concepts were developed with human superiors and human subordinates in mind.[125] Furthermore, it has been pointed out that machines have no moral agency and, by definition, it is impossible to 'punish' a LAWs.[126] Consequently, it is unreasonable to treat an autonomous agent as a subordinate in the ordinary meaning of the law.

Nevertheless, command responsibility would apply to the different human actors involved in developing and fielding LAWs. For example, military engineers supporting LAWs would presumably answer to a higher-ranking official who has a real possibility to punish or prevent any misconduct among subordinates. In line with this argument, the United States Department of Defense accepts in principle that a person authorizing the use of LAWs may be held to account, although another person is the operator.[127] Thus, the first requirement of command responsibility could be fulfilled in military operations involving deep reinforcement learning LAWs.

4.4 The mental element

The second requirement, concerning the mental element, is controversial. The *ad hoc* Tribunals and the ICC do not adopt the same approach to the mental element, which makes it debatable what is the correct principle under customary law.[128] To simplify the application to LAWs, I will opt out of that discussion and merely adopt the 'knew or had reason to know' standard.[129] ICTY has defined that standard as follows:

> [A commander] may possess the mens rea for command responsibility where: (1) he had actual knowledge, established through direct or circumstantial evidence (…) or (2) where he had in his possession information of a nature, which at least, would put him on notice of the risk of such offences by indicating the need for additional investigation in order to ascertain whether such crimes were committed or were about to be committed by his subordinates.[130]

---

[124] *Celebici case* (Appeals Chamber judgement) ICTY-96-21 (20 February 2001) para 256.

[125] Thompson Chengeta, 'Accountability Gap: Autonomous Weapon Systems and Modes of Responsibility in International Law' (2016) 45 Denver Journal of International Law and Policy 1, 31.

[126] Chengeta (n 125) 32.

[127] Michael N Schmitt, 'Autonomous Weapon Systems and International Humanitarian Law: A Reply to the Critics' (2013) Harvard National Security Journal.

[128] Cassese (n 123) 190; Cryer and others (n 95) 394.

[129] Compare Guénaël Mettraux, *The Law of Command Responsibility* (Oxford University Press 2009) ch 10.1.2.1 and API art 86(2).

[130] *Celebici case* (Trial Chamber judgement) ICTY-96-21 (16 November 1998) para 383.

Direct evidence or *indicia* could, as the quote suggests, prove actual knowledge of subordinates' criminal behaviour. Examples of relevant circumstances are the number of illegal acts, geographical location, types of troops involved and for how long time the offences occurred.[131] General awareness of some form of unlawful conduct is not sufficient to establish actual knowledge,[132] but could be a relevant factor when assessing what the superior 'had reason to know.'[133]

What the superior 'had reason to know' (constructive knowledge) depends on the specific circumstances ruling at the time.[134] The ICTY has stressed that superior responsibility is not a form of strict liability,[135] and it has rejected a negligence standard.[136] The ICTY has ruled that certain information triggers a superior's duty to investigate and examples of triggering factors are reports of breaches, a subordinate's criminal history, the subordinate's level of training, and tactical circumstances.[137] Nevertheless, a failure to obtain relevant information is not enough to presume that the superior had reason to know.[138] It must be proven that the superior deliberately refrained from using investigative means that were available to him or her.[139]

Concerning LAWs, investigating actual knowledge presents similar challenges as those discussed in *section 3.4* above. *Indicia* may be inconclusive, and a responsible superior can argue that he or she had no practical information about the convoluted logic of the LAWs, or that the agent's behaviour changed unforeseeably after interacting with the operative environment. Although in most instances that may be a strong argument to exclude a duty to investigate, the 'had reason to know' standard is more demanding and may require investigative measures. For instance, if the programmers provide a report on the performance of a LAWs that indicates a high error rate in previous missions, the *mens rea* of the commander could be presumed although the commander did not read the report. Even if the commander did read the report, but failed to understand its content, the commander would be expected to consult technological advisors available. Once the

---

[131] *Celebici case* (Appeals Chamber judgement) (n 124) para 238.
[132] *Oric case* (Appeals Chamber judgement) ICTY-03-68 (3 July 2008) paras 169-174.
[133] *Strugar case* (Appeals Chamber judgement) ICTY-01-42 (17 July 2008) para 301.
[134] *Ibid.* para 298.
[135] *Celebici case* (Appeals Chamber judgement) (n 124) paras 226, 239.
[136] *Ibid.* para 226.
[137] *Krnojelac case* (Appeals Chamber judgement) ICTY-97-25 (17 September 2003) paras 154-155.
[138] *Celibici case* (Appeals Chamber judgement) (n 124) para 226.
[139] *Ibid.*

commander has been put on investigative notice, a failure to uncover accessible information would not exclude liability.

4.5 Necessary and reasonable measures

Lastly, the third criterion of command responsibility requires a failure to take necessary and reasonable measures to prevent or punish a subordinate's crime. 'Necessary' and 'reasonable' are circumstantial standards,[140] and a superior should only be held criminally responsible for failing to take actions that are materially possible.[141] No one can be obliged to perform the impossible, but lack of formal legal competence does not *per se* exclude material possibility.[142] The commentary to API suggests that a commanding officer should act like an 'investigating magistrate' by informing superior officers, drafting incident reports, exercising disciplinary power, and remitting the case to judicial authorities.[143] However, it should be stressed that a commander cannot make up for a failure to prevent grave breaches by punishing the subordinates afterwards.[144]

It is unclear what this entails for commanders relying on deep reinforcement learning LAWs, and there is no case law since the problem is hypothetical. However, there is an ongoing discussion on command responsibility in cyber operations that could be helpful to consult. The Tallinn manual 2.0 on cyber warfare addresses command responsibility and it reasons that superiors are entitled to rely on the technical expertise of subordinates.[145] That does not, according to the manual, exclude the possibility that a commander may wilfully or negligently fail to acquire the necessary information.[146] The commander must at least be able to uphold his or her legal duty to suppress the commission of cyber war crimes, according to the manual.[147]

Applying this logic analogously, a commander responsible for deep reinforcement learning LAWs may rely on the technical expertise of his subordinates when investigating a suspected incident. If a technological advisor gives a credible explanation of the events,

---

[140] *Blaskic case* (Appeals Chamber judgement) ICTY-95-14 (29 July 2004) paras 72, 417.
[141] *Celebici case* (Trial Chamber judgement) (n 130) para 395.
[142] *Ibid.*
[143] Pilloud and others (n 119) para 3562.
[144] *Blaskic case* (Trial Chamber judgement) ICTY-95-14 (3 March 2000), para 336.
[145] Michael N Schmitt (ed), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (Cambridge University Press 2017), Rule 85(10). N.B. This is not a treaty, but the conclusions of an expert group analysing international law in cyberspace.
[146] *Ibid.*
[147] *Ibid.*

the commander is entitled to trust such a statement without being held criminally liable if a similar incident reoccurs. However, the superior may not uncritically believe a subordinate in a manner that is negligent or could be considered wilful blindness.

To define negligent behaviour *vis* machine learning LAWs, Professor Peter Margulies has suggested a threefold standard called 'dynamic diligence.'[148] Firstly, Margulies argues that a commander must make sure that the command structure contains persons who have specialized knowledge of LAWs and a separate LAWs commander could be required.[149] Secondly, the dynamic diligence standard requires frequent periodic review of the LAWs' performance in the field and regular updates of the input data as well as the algorithm.[150] Thirdly, Margulies encourages an approach to the programming of LAWs that favours interpretability of the algorithm and limits how LAWs practically could be used. For instance, the operation of a specific agent could be limited in time and distance.[151]

Margulies unmistakeably suggests a *de lege ferenda* definition of reasonable and necessary measures. If the 'dynamic diligence' standard has been properly implemented before any suspected grave breach, the responsible commander could rely upon an entire investigative infrastructure to fulfil his or her investigative duty. Competent officials could consult the periodic reviews and read code that has been consciously developed to facilitate human interpretation. In the future, Margulies' idea of dynamic diligence may inform how tribunals adjudicate superiors' duty to investigate suspected breaches. Failure to live up to that standard may, in such a future, serve as grounds for criminal liability. However, it must be stressed that it is a hypothetical standard. It is possible that Margulies is too optimistic regarding humans' ability to produce interpretable code and he fails to consider the black-box issue. No commander can be obliged to perform the impossible, and in the context of deep reinforcement learning LAWs, Margulies' dynamic diligence may be just that: Impossible.

---

[148] Peter Margulies, 'Making Autonomous Weapons Accountable: Command Responsibility for Computer-Guided Lethal Force in Armed Conflicts' in Jens Ohlin (ed), *Research Handbook on Remote Warfare* (Edward Elgar Publishing 2017).

[149] *Ibid.* 429.

[150] *Ibid.* 434.

[151] *Ibid.* 437.

5. Standards to apply when investigating

5.1 Introduction

The GC's and API do not explicitly regulate investigative standards in domestic procedures. Some procedural issues are regulated in the minimum standard protecting prisoners of war in art 105 of GCIII, which has been expanded to apply generally regardless of prisoner-of-war status.[152] Notwithstanding that provision, there is a lack of IHL norms that explicitly addresses how to investigate and prosecute war crimes.[153]

Instead, one has to rely on legal standards derived from both IHL and IHRL.[154] Central notions, sometimes referred to as 'general principles,'[155] are independence, impartiality, thoroughness, promptness, and effectiveness.[156] Moreover, a principle of transparency should be discussed in this context.[157] These principles could facilitate the materialization of the duty to investigate and clarify how to scrutinize deep reinforcement learning LAWs.

5.2 Independence and impartiality

Although independence and impartiality are two separate principles, they are intertwined to the extent that it is virtually impossible to deal with them unconnectedly. The first concept, independence, means institutional detachment from the persons allegedly implicated in the incident investigated.[158] The European Court of the Human Rights

[152] Pictet (n 107) 595 and GCIV art 146 (n 7).

[153] Amichai Cohen and Yuval Shany, 'Beyond the Grave Breaches Regime: The Duty to Investigate Alleged Violations of International Law Governing Armed Conflicts' in Michael N Schmitt and Louise Arimatsu (eds), *Yearbook of International Humanitarian Law 2011 - Volume 14* (TMC Asser Press 2012) 56; Schmitt, 'Investigating Violations of International Law in Armed Conflict' (n 19) 597.

[154] For a discussion on the interplay between the two bodies of law, see *section 1.3*.

[155] UN Human Rights Council, 'Human Rights in Palestine and Other Occupied Arab Territories', 25 September 2009, UNGA A/HRC/12/48, para 1814.

[156] UN Human Rights Council, 'Report of the Committee of independent experts in international humanitarian and human rights laws to monitor and assess any domestic, legal or other proceedings undertaken by both the Government of Israel and the Palestinian side, in the light of General Assembly resolution 64/254, including the independence, effectiveness, genuineness of these investigations and their conformity with international standards' (23 September 2010) UNGA A/HRC/15/50 para 21.

[157] UN Human Rights Council, 'Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, Philip Alston' (28 May 2010) UNGA A/HRC/14/24/Add.6 paras 87–92.

[158] UN Human Rights Council (n 156) para 22.

(ECHR) has ruled that this entails functional independence while conducting the investigation and a lack of hierarchical connection to the persons involved.[159] Regarding the second concept, impartiality, the investigative body must not be tainted with preconceived ideas of guilt or innocence.[160] It may seem quite straight-forward. However, in the context of military investigations, the dual role of military commanders as a law enforcer and a legal subject makes the independence and impartiality requirement more complex.[161] Potential factors hampering investigations into grave breaches could be a commander's fear of career repercussions, loyalty to fellow soldiers, and personal guilt.[162]

Looking at State practice, military investigations generally engage a mixture of military officers, civilian judges, legal counsel of the armed forces and independent lawyers.[163] A common argument is that military investigations must be separated from the regular military command structure and that the military command should not be able to interfere at any stage of the process.[164] A contrary position is that the chain of command is irrelevant, provided that there are procedural safeguards that guarantee an investigation free from undue influence and biases.[165] This viewpoint stresses the importance of prohibiting wrongful interference, whereas mandating a particular organizational responsibility is only of secondary interest.[166]

In relation to deep reinforcement learning LAWs, there are few entirely novel issues regarding independence and impartiality. If there is a LAWs commander, as suggested by Margulies, that person faces similar problems as any other commander investigating a potential breach. Although advanced technology may limit the number of possible investigators, it is not impossible to assemble an independent group of investigators. For example, investigative bodies could set up a permanent task force consisting of AI experts who act outside the chain of command. Such a group could include military personnel, but

---

[159] *Al Skeini v UK,* (Grand Chamber Judgment) ECHR app. No. 55721/07 (7 July 2011) para 167.

[160] UN Human Rights Council (n 156) para 23.

[161] Cohen and Shany (n 153) 65.

[162] Eugene R Fidell, *Military Justice: A Very Short Introduction* (Oxford University Press 2016) 86.

[163] Cohen and Shany (n 153) ch 2.4.1; Schmitt, 'Investigating Violations of International Law in Armed Conflict' (n 19) ch 12.3.

[164] Peter Rowe, 'How Well Do International Human Right Bodies Understand Military Courts?' in Alison Duxbury and Matthew Groves (eds), *Military justice in the modern age* (Cambridge University Press 2016) 24–25.

[165] Schmitt, 'Investigating Violations of International Law in Armed Conflict' (n 19) 628.

[166] *Ibid.*

it is essential that members of the armed forces do not investigate incidents in which they themselves are implicated.

5.3 Effectiveness and thoroughness

This standard calls for a complete and comprehensive investigation.[167] In peacetime, one could argue that responsible authorities must undertake autopsies and medical examinations, interview all relevant witnesses, visit the scene of the crime, and more depending on the case at hand.[168] In times of armed conflict, however, it is impossible to investigate every death or reach the same level of effectiveness and thoroughness.[169] The right to life is a non-derogable standard, and procedural guarantees including investigative standards must uphold this right, yet what reasonably could be done is limited in case of armed conflict.[170]

As pointed out in *section 1.3,* evidence may have been destroyed in battle, travel is usually risky, and standard forensic tools may be unavailable. Nevertheless, it is conceivable that a party to an armed conflict may have a high degree of control over a specific area and in that area have the capability to conduct an investigation on par with peacetime standard. Therefore, it has been argued that accepting a less effective investigation should only be done on a case-by-case basis and only in response to specific constraints.[171] Furthermore, a commander's duty to punish IHL violations include an obligation to conduct effective investigations into alleged war crimes,[172] and effective response requires that the commander takes reasonable measures to establish the facts.[173] What is reasonable depends on the circumstances.[174]

Is it possible to effectively and thoroughly investigate the black box that is deep reinforcement learning LAWs? There are reasons to be doubtful. Apart from the usual issues of investigating incidents in armed conflict, the technology ads a layer of

---

[167] UN Human Rights Council (n 156) para 24.

[168] *Ibid.*

[169] *Ibid.* para 32.

[170] Commission on Human Rights, 'Civil and political rights, including the question of disappearances and summary executions – Extrajudicial, summary or arbitrary executions – Report of the Special Rapporteur, Philip Alston' (8 March 2006) Economic and Social Council E/CN.4/2006/53 para 36.

[171] *Ibid.*

[172] *Boskoski and Tarculovski case*, (Trial Chamber Judgement) ICTY-04-83 (10 July 2008) para 418.

[173] *Strugar case,* (Trial Chamber Judgement) ICTY-01-42 (31 January 2005) para 376.

[174] See *section 4.5*

uninterpretable code. An investigator trying to establish the facts would in a best-case scenario have to explain an extensive audit trail of hidden neuron-units. In a worst-case scenario, there would not even be an audit trail.

However, researchers are trying to make that task manageable by applying more AI. The United States' Defense Advanced Research Projects Agency is currently funding a research project on how to make deep learning algorithms explain themselves by applying a separate deep learning algorithm.[175] The theory is that an 'investigation algorithm' could provide information about the 'operative algorithm' in a way that is interpretable for humans and thereby open up the black box. In case that algorithm is in place, it is conceivable that States could conduct effective investigations of deep reinforcement learning LAWs.

5.4 Promptness

In the GC's, there is a maximum limit to the promptness requirement. After giving notice, the accused must have one week to choose legal counsel, and then the defence must be given two weeks to prepare its case.[176] Regarding minimum requirement, there is no definite time limit. The ECHR has ruled that there is an implicit requirement of reasonable expedition in the context of a State's duty to conduct effective investigations, but the court interpreted that requirement circumstantially.[177]

In some cases, the nature of the crime may call for an especially prompt investigation. Immediate action is usually required concerning an on-going crime where an individual's life is in danger or concerning an incident where evidence is likely to disappear.[178] Examples thereof are reoccurring attacks on the civilian population and instances of torture or extrajudicial killings conducted by State forces. In other cases, however, a delay of an investigation might allow for a more comprehensive inquiry. At the end of hostilities in a specific area, some of the constraint on investigative possibilities tend to lose relevance.[179]

The nature of deep reinforcement learning LAWs could require special promptness in investigations. Surely, if the involved LAWs is still in use, a swift inquiry is necessary to

[175] 'Research Aims to Make Artificial Intelligence Explain Itself' (*Life at OSU*, 5 June 2017) <https://today.oregonstate.edu/archives/2017/jun/research-aims-make-artificial-intelligence-explain-itself> accessed 27 November 2018.

[176] Pictet (n 107) 595, GCIV art 105 (n 7) and GCIV art 146 (n 7).

[177] *Isayeva, Yusupova and Bazayeva v. Russia*, (n 105) paras 209–213.

[178] UN Human rights council (n 156) para 25.

[179] *Ibid.* para 32.

prevent further breaches. Moreover, if a State fails to take investigative action within a reasonable time frame, essential parts of the evidence may be lost since software is continuously updated through reinforcement learning and records of input data can be deleted. There is no explicit requirement to log information about the LAWs, but if there are such logs, programmers could presumably erase them. Because of these characteristics, an especially prompt investigation is called upon to secure evidence and prevent further breaches. However, assessing the complete damage of LAWs may be impossible during an on-going conflict, and the promptness requirement may be moderate in this regard.

5.5 Transparency

National security concerns may exclude transparency in investigations of violations taking place during an armed conflict.[180] In peacetime, the public's right to be informed is protected, and transparency is seen as good governance.[181] In wartime, on the other hand, there is no IHL document directly promoting transparency, and most IHRL treaties include an exemption concerning national security.[182] Nevertheless, special UN rapporteur on extra-judicial killings, Philip Alston, has stated that the principle of transparency can be derived from common article 1 of the GC's, and the grave breaches regime.[183] A lack of transparency would in practice give States a virtual and impermissible license to kill, Alston argued,[184] and he has further reasoned that any duty to investigate would be void if State parties did not have to demonstrate that they complied with such a duty.[185]

There is no shortage of counterarguments that generally support a principle of secrecy. Concealing information may support ruses of war, which is an explicitly legal tactic.[186] Military necessity may be invoked in general, but there is no specific formula that could be superimposed in all circumstances.[187] Professor Schmitt has argued that secrecy in investigations is called upon to protect military tactics, cooperative witnesses, and classified

[180] *Ibid.*
[181] Ben-Naftali and Peled (n 103) 321.
[182] *Ibid.* 322.
[183] UN Human Rights Council (n 157) para 88.
[184] *Ibid.*
[185] Philip Alston, 'The CIA and Targeted Killings beyond Borders' (2011) 2 Harvard National Security Journal 283, 311.
[186] API art 37(2); Hague Convention (IV) respecting the Laws and Customs of War on Land and its annex: Regulations concerning the Laws and Customs of War on Land. The Hague, 18 October 1907, art 24.
[187] Ben-Naftali and Peled (n 103) 345.

weapons data.[188] A brute example of invoking military necessity to hide the result of an investigation is when the United States claimed that they could not disclose their investigation into the Abu Ghraib prison in Iraq because the reaction could jeopardise the security of military personnel posted in Iraq.[189]

In response, attempts have been made to limit the scope of military necessity. For instance, special UN rapporteur Emmanuel Decaux, presented a principle that an independent monitoring body should review any argument concerning military secrecy.[190] Nevertheless, the relationship between military necessity and transparency remains mostly undefined. Specific legal requirements are scarce, and States' position on the matter is often the result of policy considerations rather than legal considerations.[191]

In theory, it is possible that deep reinforcement learning LAWs could enhance transparency. As explained in *section 2*, the system must receive some kind of input. Depending on the sensors of the LAWs, that input could be optical images, infrared thermography, radar signals, sounds, and other sources of information that are physically detectable. If the LAWs recorded all its input, it would be possible to be fully transparent about what happened by simply providing investigators and the general public access to the recorded material.[192]

Although the technology is capable of transparency, there is a lack of legal requirements enforcing public disclosure of investigations. Transparency is a principle often invoked concerning investigations of grave breaches, but seldomly defined legally. The secrecy surrounding national security gives States an extensive right to determine what is made public and what is kept secret. Hence, an investigator could merely say that the footage from a LAWs has been reviewed and that no misconduct is suspected, without actually disclosing the video.

---

[188] Schmitt, 'Investigating Violations of International Law in Armed Conflict' (n 19) 629.

[189] Ben-Naftali and Peled (n 103) 352.

[190] UN Commission on Human Rights, 'Civil and political rights, including the question of independence of the judiciary, administration of justice, impunity – Issue of the administration of justice through military tribunals – Report submitted by the Special Rapporteur of the sub-commission on the promotion and protection of human rights, Emmanuel Decaux' (13 January 2006) Economic and Social Council E/CN.4/2006/58 principle no 10.

[191] Schmitt, 'Investigating Violations of International Law in Armed Conflict' (n 19) 629.

[192] A UN special rapporteur suggested this, see: UN Human Rights Council 'Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, Christof Heyns' UNGA A/HRC/23/47, para 81

6. Closing remarks

Returning to AlphaGo, it is worth mentioning that the DeepMind team eventually managed to explain the "weird" lazy-looking defensive moves. The algorithm did not care if it won by half a point or by 20 points, it only cared about winning. Given that logic, a low-risk low-reward move was more rational than a high-risk high-reward move.[193] The human operators could not keep up with the system in real time, but they got there eventually.

I acknowledge that the idea of some kind of AlphaGo-killing machine is abstract or even outright ridiculous. However, if the legal community wishes to keep up to speed with technological progress, it is important to allow oneself to look silly. AI is on the rise, and although we do not know precisely how it will function or how it will be regulated, we must prepare by asking the right type of questions.

As to answers, it is difficult to say anything with certainty. My position is that deep reinforcement learning LAWs is a potentially disruptive technology, and if States are serious about upholding their duty to investigate grave breaches of IHL, they should tread carefully. I don't argue in favour of or against a pre-emptive ban on LAWs; I simply conclude that significant investigative issues remain to be resolved. And, to paraphrase Kasparov, if I live to see an autonomous weapons system that is well beyond anyone's understanding, I will be scared.

[193] Kohs (n 2) at 79 mins

7. List of Authorities

7.1 Doctrine

Alston P, 'The CIA and Targeted Killings beyond Borders' (2011) 2 Harvard National Security Journal 283

Ben-Naftali O and Peled R, 'How Much Secrecy Does Warfare Need?' in Andrea Bianchi and Anne Peters (eds), *Transparency in International Law* (Cambridge University Press 2013)

Boulanin V and Verbruggen M, *Mapping the Development of Autonomy in Weapon Systems*, (Stockholm International Peace Research Institute 2017)

Carlsen H and others, 'Assessing Socially Disruptive Technological Change' (2010) 32 Technology in Society 209

Cassese A, *Cassese's International Criminal Law* (3rd edn, Oxford University Press 2013)

Chengeta T, 'Accountability Gap: Autonomous Weapon Systems and Modes of Responsibility in International Law' (2016) 45 Denver Journal of International Law and Policy 1

Chengeta T, 'Defining the Emerging Notion of Meaningful Human Control in Weapon Systems' (2016) 49 New York University Journal of International Law and Politics 833

Cohen A and Shany Y, 'Beyond the Grave Breaches Regime: The Duty to Investigate Alleged Violations of International Law Governing Armed Conflicts' in Michael N Schmitt and Louise Arimatsu (eds), *Yearbook of International Humanitarian Law 2011* (vol 14, T.M.C Asser Press 2012)

Cryer R and others, *An Introduction to International Criminal Law and Procedure* (2nd edn, Cambridge University Press 2010)

Diesen C, *Bevis 7: Bevisprövning i förvaltningsmål* (Norstedts Juridik AB 2003)

Fidell ER, *Military Justice: A Very Short Introduction* (Oxford University Press 2016)

Goodfellow I, Bengio Y and Courville A, *Deep Learning* (The MIT Press 2016)

Jackson D, 'Reporting and Investigation of Possible, Suspected, or Alleged Violations of the Law of War' (2010) Army Lawyer

Johansson L, *Äkta robotar* (Fri Tanke Förlag 2015)

Kleffner JK, *National Suppression of Core Crimes* (Oxford University Press 2008)

Margulies P, 'Making Autonomous Weapons Accountable: Command Responsibility for Computer-Guided Lethal Force in Armed Conflicts' in Jens Ohlin (ed), *Research Handbook on Remote Warfare* (Edward Elgar Publishing 2017)

Matthias A, 'The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata' (2004) Ethics and Information Technology 175

McFarland T and McCormack T, 'Mind the Gap: Can Developers of Autonomous Weapons Systems Be Liable for War Crimes' (2014) 90 International Law Studies Series

Mettraux G, *The Law of Command Responsibility* (Oxford University Press 2009)

Rowe P, 'How Well Do International Human Right Bodies Understand Military Courts?' in Alison Duxbury and Matthew Groves (eds), *Military justice in the modern age* (Cambridge University Press 2016)

Schmitt MN, 'Investigating Violations of International Law in Armed Conflict' in *Essays on Law and War at the Fault Lines* (T.M.C Asser Press 2011)

Schmitt MN, 'Autonomous Weapon Systems and International Humanitarian Law: A Reply to the Critics' (2013) Harvard National Security Journal

Schmitt MN (ed), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (Cambridge University Press 2017)

Tan AML, 'The Duty to Investigate Alleged Violations of International Humanitarian Law: Outdated Deference to an Intentional Accountability Problem' 49 International Law and Politics 182

7.2 State Practice

Statement by the United States at the 2014 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS), May 16, 2014 (as cited in Chengeta (n 29))

Swedish government regulation re disciplinary responsibility [*Förordning (1995:241) om disciplinansvar inom totalförsvaret m.m.*]

Swedish law re disciplinary misconduct [*Lag (1994:1811) om disciplinansvar inom totalförsvaret, m.m*]

Swedish preparatory works re criminal investigation [*Prop 1994/95:23 Ett effektivare brottmålsförfarande*]

U.S. Department of Defense, 'Directive 3000.09, Autonomy in Weapon Systems'

7.3 International organizations

7.3.1 ICRC

ICRC, *Commentaries on the first Geneva Convention* (Cambridge University Press 2016)

Pictet JS (ed), *Commentary: Fourth Geneva Convention Relative to the Protection of Civilian Persons in Time of War* (ICRC 1958)

Pilloud C and others, *Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949* (Kluwer Academic Publishers 1987)

7.3.2 United Nations

Commission on Human Rights, 'Civil and political rights, including the question of disappearances and summary executions – Extrajudicial, summary or arbitrary executions – Report of the Special Rapporteur, Philip Alston' (8 March 2006) Economic and Social Council E/CN.4/2006/53

Commission on Human Rights, 'Civil and political rights, including the question of independence of the judiciary, administration of justice, impunity – Issue of the administration of justice through military tribunals – Report submitted by the Special Rapporteur of the sub-commission on the promotion and protection of human rights, Emmanuel Decaux' (13 January 2006) Economic and Social Council E/CN.4/2006/58

Human Rights Council, 'Human Rights in Palestine and Other Occupied Arab Territories' (25 September 2009) UNGA A/HRC/12/48

Human Rights Council, 'Report of the Committee of independent experts in international humanitarian and human rights laws to monitor and assess any domestic, legal or other proceedings undertaken by both the Government of Israel and the Palestinian side, in the light of General Assembly resolution 64/254, including the independence, effectiveness, genuineness of these investigations and their conformity with international standards' (23 September 2010) UNGA A/HRC/15/50

Human Rights Council, 'Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, Christof Heyns' (9 April 2013) UNGA A/HRC/23/47

Human Rights Council, 'Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, Philip Alston' (28 May 2010) UNGA A/HRC/14/24/Add.6

International Law Commission, finalized Koskenniemi M, 'Fragmentation of international law: Difficulties arising from the diversification and expansion of international law' (1 May-9 June and 3 July-11 August 2006) UNGA A/CN.4/L.682

Rome Statute of the International Criminal Court (17 July 1998) UN Doc A/CONF.183/9

7.4 Jurisprudence

7.4.1 European Court of Human Rights

Al Skeini v UK, Grand Chamber judgment of 7 July 2011, application no 55721/07

Isayeva, Yusupova and Bazayeva v. Russia, Judgement of 24 February 2005, application nos. 57947/00, 57948/00, 57949/00)

7.4.2 International Court of Justice

Legal Consequences of the Construction of a Wall in the Occupied Palestinian Territory, Advisory Opinion, 2004, ICJ Rep 136

Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion, 8 July 1996, ICJ Rep 226.

7.4.3 International Criminal Tribunal for the former Yoguslavia

Blaskic, Prosecutor v. (Case No. IT-95-14)

— Judgement, Trial Chamber, 3 March 2000.

— Judgement, Appeals Chamber, 29 July 2004

Boskoski and Tarculovski, Prosecutor v. (Case No. IT-04-83)

— Judgement, Trial Chamber, 10 July 2008

Celebici case: Prosecutor v. Delalic, Mucic, Delic, and Landzo (Case No. IT-96-21)

— Decision on the motions by the Prosecution for protective measures for the prosecution witnesses pseudonymed 'B' through to 'M', Trial Chamber, 28 April 1997

— Judgement, Trial Chamber, 16 November 1998

— Judgement, Appeals Chamber, 20 February 2001

Hadzihasanovic and others, Prosecutor v. (Case No. IT-01-47)

— Judgement, Trial Chamber, 15 March 2006

Krnojelac, Prosecutor v. (Case No. IT-97-25)

— Judgement, Appeals Chamber, 17 September 2003

Kunarac, Prosecutor v. (Case No. IT-96-23& IT-96-23/1)

— Judgement, Appeals Chamber, 12 June 2002

Oric, Prosecutor v. (Case No. IT-03-68)

— Judgement, Appeals Chamber, 3 July 2008

Strugar, Prosecutor v. (Case No. IT-01-42)

— Judgement, Trial Chamber, 31 January 2005

— Judgement, Appeals Chamber, 17 July 2008

7.4.4 International Criminal Court

Bemba Gombo, Prosecutor v. (Case No. ICC-01/05-01/08)

— Decision Pursuant to Article 61(7)(a) and (b) of the Rome Statute on the Charges of the Prosecutor Against Jean-Pierre Bemba Gombo, Pre-Trial Chamber, 15 June 2009

7.4.5 Other military tribunals

Röchling, French Government commissioner v. (Superior Military Government Court of the French occupation zone in Germany)

— Judgement, 30 June 1948, as reported in *Trials of War Criminals before the Nuremberg Military Tribunals Under Control Council Law No. 10,* vol 14 p. 1119.

Zyklon B case: Bruno Tesch, Joachim Drosihn and Karl Weinbacher; Prosecutor v. (British Military Court case no. 9)

— Judgement, 8 March 1946, as reported in *I Law reports of Trial of War criminals 93, 94* p. 103.

7.5 News articles and online sources

'A Beginner's Guide to Deep Reinforcement Learning' (*Skymind*) <http://skymind.ai/wiki/deep-reinforcement-learning> accessed 29 November 2018

'AlphaGo' (*DeepMind*) <https://deepmind.com/research/alphago/> accessed 6 December 2018

'Automatic Target Recognition of Personnel and Vehicles from an Unmanned Aerial System Using Learning Algorithms *(Small Business Innovation Research (SBIR) program)* <www.sbir.gov/sbirsearch/detail/1413823> accessed 4 October 2018

Bleicher A, 'Demystifying the Black Box That Is AI' (Scientific American) < www.scientificamerican.com/article/demystifying-the-black-box-that-is-ai/> accessed 27 November 2018

'Campaign to Stop Killer Robots' <www.stopkillerrobots.org/> accessed 19 December 2018

'Deep Reinforcement Learning' (*DeepMind*) <https://deepmind.com/blog/deep-reinforcement-learning/> accessed 30 November 2018

'Definition of BLACK BOX' (*Merriam Webster)* < www.merriam-webster.com/dictionary/black+box> accessed 29 November 2018

'Google Ditches Department of Defense, Updates Its Code of Ethics' (Futurism) <https://futurism.com/maven-google-military-tech> accessed 30 November 2018

Kohs G, AlphaGo [Documentary] (2017) <www.netflix.com/title/80190844> accessed 15 November 2018

Krauthammer C, 'Be Afraid' The Weekly Standard <www.weeklystandard.com/charles-krauthammer/be-afraid-9802> accessed 1 November 2018

Li Y, 'Deep Reinforcement Learning: An Overview' [2017] arXiv:1701.07274 [cs] <http://arxiv.org/abs/1701.07274> accessed 30 November 2018

Mnih V and others, 'Playing Atari with Deep Reinforcement Learning' <https://arxiv.org/abs/1312.5602> accessed 29 November 2018

'Research Aims to Make Artificial Intelligence Explain Itself' (*Life at OSU*, 5 June 2017) <https://today.oregonstate.edu/archives/2017/jun/research-aims-make-artificial-intelligence-explain-itself> accessed 27 November 2018

Silver D, 'Deep Reinforcement Learning' <http://videolectures.net/rldm2015_silver_reinforcement_learning/> accessed 30 November 2018

Strachan ALS, *Can We Build a Brain? [Documentary]* (2018) <www.svtplay.se/video/19065271/vetenskapens-varld/vetenskapens-varld-de-smarta-maskinernas-tid> accessed 10 December 2018

Zahavy T, Zrihem NB, and Mannor S, 'Graying the Black Box: Understanding DQNs' <https://arxiv.org/abs/1602.02658> accessed 19 December 2018