Algorithmic fog of war: When lack of transparency violates the law of armed conflict

Kwik, J.; Van Engers, T.

DOI
10.3233/FRL-200019

Publication date
2021

Document Version
Final published version

Published in
Journal of Future Robot Life

Journal of Future Robot Life 2 (2021) 43–66
DOI 10.3233/FRL-200019
IOS Press

Algorithmic fog of war: When lack of transparency violates the law of armed conflict

Jonathan Kwik * and Tom Van Engers
Faculty of Law, University of Amsterdam, Amsterdam 1001NA, Netherlands

Abstract. Under international law, weapon capabilities and their use are regulated by legal requirements set by International Humanitarian Law (IHL). Currently, there are strong military incentives to equip capabilities with increasingly advanced artificial intelligence (AI), which include opaque (less transparent) models. As opaque models sacrifice transparency for performance, it is necessary to examine whether their use remains in conformity with IHL obligations. First, we demonstrate that the incentives for automation drive AI toward complex task areas and dynamic and unstructured environments, which in turn necessitates resort to more opaque solutions. We subsequently discuss the ramifications of opaque models for foreseeability and explainability. Then, we analyse their impact on IHL requirements from a development, pre-deployment and post-deployment perspective. We find that while IHL does not regulate opaque AI directly, the lack of foreseeability and explainability frustrates the fulfilment of key IHL requirements to the extent that the use of fully opaque AI could violate international law. States are urged to implement interpretability during development and seriously consider the challenging complication of determining the appropriate balance between transparency and performance in their capabilities.

Keywords: Transparency, interpretability, foreseeability, weapon, International Humanitarian Law, armed conflict, autonomy, autonomous weapon

1. INTRODUCTION

Machine learning has been touted as a great step forward for the versatility and general applicability of artificial intelligence (AI) (Cukier, 2018). Continued advancements in machine learning models have made them increasingly ubiquitous in optimising performance for more nuanced tasks and complicated working environments (Doshi-Velez and Kim, 2017). The same trend is true in the military domain, where automation has been identified as a powerful force multiplier, extending human capability and improving both speed and accuracy (Department of Defense, 2012; Rosenberg and Markoff, 2016). It has been described as the "hidden, invisible power" of modern weapons (Ferguson, 2001, at 105). Automation is utilised to great effect for intelligence, surveillance and reconnaissance (ISR), military decision-making, and taking over functions such as verification or targeting (Parakilas and Bryce, 2018; ICRC, 2019; Abaimov and Martellini, 2020).

As in the civil sector, many of these advancements were made possible by the development of data-driven algorithms, which opened avenues for the automation of functions and tasks which would previously have been extremely challenging to program. As these algorithms enable increasingly high-risk functions involving the safety of civilians, however, concerns have been raised with regard to the lack of explainability and interpretability which often comes with the use of machine learning techniques (Biran and Cotton, 2017; Mueller et al., 2019; Defense Innovation Board, 2019). Authors have written about the value of transparency from many different perspectives, such as improving user trust (Saariluoma, 2015; Ribeiro et al., 2016), preventing bias (Doshi-Velez and Kim, 2017;

*Corresponding author. E-mail: h.c.j.kwik@uva.nl; Tel.: +31653110578.

ICRC, 2019), debugging (Molnar, 2019; Miller, 2019) and improving knowledge (Lombrozo, 2006; Wilkenfeld and Lombrozo, 2015).

The use of more opaque models for military purposes, however, raises additional considerations in light of the requirements set forth in international norms. In particular, International Humanitarian Law (IHL), often referred to as the Law of Armed Conflict, regulates the conduct of belligerents to safeguard the interests of parties not actively involved in the conflict and to prevent the infliction of harm which is not necessary to achieve a military victory (Kalshoven and Zegveld, 2001; United Kingdom, 2004). IHL regulates both the use of capabilities empowered by AI and sets requirements for the capabilities themselves.

Even disregarding transparency concerns, developers and users of military capabilities already grapple with challenges that emerge from the use of increasingly complicated AI, such as understanding their different way of thinking (Statler, 1993; Department of Defense, 2015; Bryce and Parakilas, 2018), how the algorithm learns (Russell and Norvig, 2010; ICRC, 2019; Etzioni and Etzioni, 2016), and how this affects the chain of responsibility (Crootof, 2015b; Keeley, 2015). Due to the prevalence of concepts in IHL that depend on a certain notion of foreseeability and predictability, the lack of interpretability creates additional obstacles to the proper implementation of obligations under IHL. which must be addressed.

In this article, we discuss the ramifications of the use of opaque (untransparent) AI in weapon capabilities vis-à-vis legal requirements for their lawful use from an IHL perspective. We argue that a shift to less transparent AI is a corollary of the military needs and circumstances to which such models can provide a solution. At the same time, this shift requires due consideration during development, pre-deployment and post-deployment phases to retain congruency with IHL exigencies. To benefit optimally from AI technology, performance must be maximised, but only to the extent that the product remains within the constraints of the law.

This article is structured as follows. First, an overview is provided of the military, political and economic incentives that drive automation in weapon capabilities. This is followed by a general taxonomy on the tasks these algorithms are likely to fulfil and in what environments they will operate, how these factors necessitate the shift to more opaque AI, and what consequences and challenges emerge from such use. Then, three categories of legal requirements which contain provisions most affected by these challenges are discussed: obligations during development, deployment, and post-deployment. We conclude that fully opaque AI that offer no degree of interpretability are inconsistent with State obligations under IHL, and recommend a proper balance be developed between the benefits and opportunity costs of implementing transparency.

We primarily discuss capabilities in the sense of robots, i.e., "physical agents that perform tasks by manipulating the physical world" (Russell and Norvig, 2010, at 971), and not cyberweapons or digital-only systems developed to support or replace tactical-level decisions. In addition, while transparency issues indisputably pose problems for many legal domains such as criminal law, civil law, human rights law and tort law, the legal analysis in this paper focuses uniquely on legal obligations stemming from IHL, except where resort to a specific branch of law is mandated (e.g. legal accountability).

2. INCENTIVES FOR AUTOMATION

Vladimir Putin has famously said that "the one who becomes the leader in this sphere [AI] will be the ruler of the world" (Dailymail, 2017). Major military powers seem to subscribe to this sentiment,

and have placed emphasis on the development and incorporation of AI into their capabilities (Department of Defense, 2012; Wilson, 2020). According to Bowne (2019, at 75), adopting AI is crucial to "maintain a technological advantage over military capabilities of near-peer competitors". The US Department of Defence (DoD) (Rosenberg and Markoff, 2016) has justified their large investments in autonomous technology as necessary to maintain its "military edge" over rivals such as China and Russia, and to match the pace of its allies, such as the United Kingdom and Israel.

Previously, a large focus was placed on the efficacy of unmanned (but still remote-controlled) systems such as drones or unmanned ground vehicles for extending the reach of the individual soldier, allowing him to see and reach further, and enabling consistent surveillance and action (Arkin, 2013; Crootof, 2015a). An important political incentive of unmanned systems was the removal of the soldier's physical presence on the battlefield (Heyns, 2013): unmanned systems allow the projection of power without vulnerability, making them risk-free forms of warfare that do not entail internal public accountability for the lives of nationals. More succinctly, Singer (2009, at 31) reported a US Navy officer expressing that "when a robot dies, you don't have to write a letter to its mother".

In contrast, remotely-controlled systems retain the significant disadvantage of processing time, which cannot go beyond what is humanly possible. In military circles, processing time is often presented in terms of John Boyd's (1996) OODA loop (Observe-Orient-Decide-Act). Optimising the time to complete this process – or at least completing it faster than the opponent – would be the key to victory (Osinga, 2005; Marra and McNeil, 2013; Brown, 2018). The OODA loop is largely analogous to the model of human information processing proposed by Parasuraman et al. (2000), which comprises information Acquisition, Analysis, Decision and Action. Completing the feedback loop is accelerated by delegating some or all steps in the loop to an AI which is granted sufficient rights to execute its function without requiring human authorisation. An air traffic controller can, for example, work faster if an algorithm pre-sorts and highlights landing priorities of incoming aircraft, a task a human cannot perform with the same speed as an optimised algorithm can. This demonstrates that automation in even a single phase (here Acquisition) can improve the overall completion speed of the task.

An improvement in processing speed is commonly cited as the most important benefit of implementing automation from a military perspective (Heyns, 2013; Scharre, 2015, 2016). In addition, the competitive aspect of the OODA loop theory entails that to maintain military efficiency vis-à-vis an adversary, functions and tasks that have been automated by the adversary also will have to be automated in one's own capabilities, as human processing simply would not be able to 'keep up' with the rate set by the opponent. Some authors refer to this phenomenon as the 'increased battlefield tempo' brought about by technology (Arkin, 2013; Sparrow, 2016; Wilson, 2020). For example, Thurnher (2012, at 80) speculates that "future battles will likely occur at such a high tempo that human controllers may not be able to direct drone forces to rapidly counter enemy actions". Speed is particularly crucial in reactive capabilities where the human decision-making tempo is simply insufficient, for example in counter-projectile and missile systems (Geiss, 2015; Hawley, 2007; United States Navy, 2019). Theunissen and Suarez (2015, at 189) describe these as situations where human default or error must be avoided, where "immediate action is needed to prevent a catastrophic event".

In addition to speed, another frequently-cited reason for automating functions is related to removing the requirement to maintain a communication link with the system. Mayer (2015) points out that while remote-controlled systems can also fulfil 'dirty, dull and dangerous' tasks, the true value of automation is the capacity to continue functioning in environments where communications are unavailable. Unavailability can occur naturally or due to adversarial action. Certain environments simply preclude the possibility of effectively maintaining a link at all times, such as those underwater or

underground (Anderson and Waxman, 2013; Canada, 2016). Others are communication-denied environments where the enemy is actively attacking or jamming the link with the system. Delegating certain functions to the system will remove this vulnerability if such adversarial action is foreseeable, thereby increasing utility and versatility (Crootof, 2015a; Sparrow, 2016).

A final claim that is made with regard to the benefits of automation is related to battlefield amelioration, i.e., that removing humans from (certain parts of) the decision-making loop will normatively improve the battlefield situation. Often, the improvement is framed as an indirect result of superior performance leading to a better humanitarian standard. If the delegation of certain functions to algorithms statistically results in better speed, increased accuracy and performance and lower error rates, this would lead to fewer unintended or unneeded damage or harm to civilians (Williams, 2015; Mayer, 2015). This sentiment mirrors some related discussions in the civil domain, such as with respect to autonomous cars (Yadron and Tynan, 2016). On the other hand, some sources (Sassoli, 2014; Department of Defense, 2018; Defense Innovation Board, 2019) more directly state their belief that automation can enhance compliance with IHL norms by reducing the amount of violations (vis-à-vis if human involvement is maintained). This is often supported by the argument that many human violations occur due to non-technical (non-performance) related shortcomings, such as irrationality, emotion and psychological and sociological biases (Lin et al., 2009; UK Ministry of Defence, 2011; Deng, 2015).

3. TASKS AND ENVIRONMENTS

Already in 1987, Jean de Preux (Sandoz et al., 1987, at ¶1476) in his commentary on the Additional Protocols to the Geneva Conventions predicted a future of technological developments which would lead to "the automation of the battlefield in which the soldier plays an increasingly less important role". The existence of many capabilities today which can operate with partial or even complete autonomy has proven De Preux's statement correct. While some sources are under the assumption that it concerns a future phenomenon (Human Rights Watch, 2012; Wareham, 2018), AI has been empowered for decades in various forms to improve the efficiency of capabilities, such as by managing flight controls, intercepting incoming projectiles and identifying targets (Dahm, 2012; Crootof, 2015a; Theunissen and Suarez, 2015). Bryce and Parakilas (2018) underline that AI already exists in even "mundane ways" in current-day weapon capabilities; rather, it is a question of in which functions and capacities such AI is integrated, and how these trends might develop in the foreseeable future.

The DoD (2015, at 2) asserts that the "level of complexity of the environment, the task at hand, and the broader mission all influence the level of difficulty of making effective decisions". In the context of weapon systems, there is a gradual demand for more generalisation and versatility with respect to most task areas. This is a logical necessity of the motivations driving automation in the first place, as discussed in Section 2. As a result, AI must adapt accordingly to fulfil these functions in the context of both more complicated tasks and environments.

3.1. Tasks

While increasing autonomy through AI is certainly the logical and perhaps inevitable step forward for many military technologies (Beard, 2014), authors (Scharre, 2015, 2016; Parakilas and Bryce, 2018) have also ruled out the *complete* replacement of human decision-making as improbable, at least in the near future. Delegation will continue to increase in quantity, but it will remain limited to discrete and

limited functions. As such, Boulanin (2016, at 7) points out that it is difficult to speak of autonomy as a blanket term; rather, it is a characteristic "that can be attached to a large variety of functions in weapon systems". Many authors (Williams, 2015; Arnold, 2015; Scharre, 2016) agree that it is generally more constructive to speak of autonomous *functioning* in a system instead of labelling the system itself as autonomous. To this end, a framework is helpful in distinguishing the major categories of tasks for which AI can be incorporated to improve capability performance.

Boulanin (2016) proposed a framework which describes five different "task areas" unto which automation can be implemented: *Mobility* (platform movement and navigation), *Health* (survival management, e.g. refuelling or self-repair functions), *Interoperability* (communication with other agents in the environment), *Intelligence* (analysis of tactical or strategic battlefield data) and *Force* (detection, identification and engagement of targets). The first three tasks (Mobility, Health and Interoperability) are often referred to as 'operational functions' while the latter two (Intelligence and Force) are often referred to as 'critical functions'. Boulanin accurately asserts that it is frequently the critical task areas which are deemed more problematic, or at least which raise more concerns. In the context of discussions on the legality of autonomous weapons, a frequently-cited definition by the ICRC (Davison, 2017, at 5) defines them as any weapon system with critical functions, i.e. those which "can select (search for, detect, identify, track or select) and attack (use force against, neutralize, damage or destroy) targets without human intervention". This perspective aligns with the Force task area in Boulanin's model.

In the context of analysing consequences of a lack of transparency, it is important to emphasise that a holistic analysis is necessary: frequently, it is not merely algorithms in the Force task area that are relevant. On the one hand, it is possible that problems related to opacity only concern a single task area (Keeley, 2015), such as *Why did the system move to this location?* (Mobility) or *Why was this building targeted?* (Force). On the other hand, programs in 'non-critical' task areas which manage positioning (Mobility), fuel and damage estimation (Health) and coordination and user interface (Interoperability) can influence or even be determinant of the performance of critical ones (e.g. Force), and thus must be scrutinised together if applicable. It is also very possible that different task areas may rely on the same technology or software, such as the same perception technology governing both the system's spatial navigation and target recognition functions (Boulanin, 2016).

Scherer (2016, at 363) emphasises that the "complexity and scope of tasks that will be left in the hands of AI will undoubtedly continue to increase" and AI will be required to move from more skill- and rule-based tasks toward knowledge- and goal-based ones (Wilson, 2020). The former tasks comprise simple sensory-motor actions or clear rules or subroutines. Historically, most AI integration has been centred on these tasks because they were easier to represent mathematically and complemented native human deficiencies such as boredom, complacency and cognitive difficulties in remembering rigid but extensive rules (Cummings, 2018). As the AI's roles shift toward more dynamic or complex functions such as target classification, target prioritisation and target engagement, a shift toward more knowledge-based reasoning or deep learning models that enable flexible decision-making will be required (Keeley, 2015; Abaimov and Martellini, 2020). The increased complexity of the envisaged tasks introduces a greater degree of uncertainty and predictability. In addition, if a level of randomisation is incorporated to pre-emptively counter enemies who attempt to 'game' the system by manipulating the system's percepts (Molnar, 2019), another dimension of unpredictability is added.

3.2. Environments

Environments provide percepts to agents and allow agents to perform actions and interact with them (Frank et al., 2001). The type of environment in which an agent is expected to function greatly affects

the level of sophistication its AI requires to perform well, both in a technical and legal sense (Department of Defense, 2015). Russell and Norvig (2010) proposed different dimensions through which an environment's nature can be analysed, the most relevant of which are summarised below:

- ***Certainty***. To what extent the agent's perceptors give access to the complete state of the environment relevant to the task (*observability*), and whether an action always leads to the same subsequent state in the environment (*determinism*). Environments that are not fully observable and not deterministic are regarded as *uncertain*.
- ***Population***. Whether the environment contains other agents that are influenced and can influence the environment. Moya and Tolk (2007) further subdivide multiagent environments based on the population size, diversity, whether they operate cooperatively or competitively, and whether new agents can be added.
- ***Dynamicity***. Whether the environment can change while the agent is deliberating. Static environments are those which (it can be assumed) remain unchanged while the agent is evaluating. Otherwise, it is dynamic.

Russell and Norvig (2010) comment that robots (i.e. physical agents) usually operate in environments which are partially observable, non-deterministic, multiagent, and dynamic. Tolk (2015) agrees that in general, the 'real world' is non-deterministic and contains many uncertainties, as does Wooldridge (2001, at 7): "The physical world is a highly dynamic environment." Complex battlefields, especially urban environments, contain a multitude of civilian objects and agents and dense infrastructures that must be understood and navigated (Asaro, 2006). The uncertain nature of the environment will necessitate making predictions, while its dynamic nature will place pressure on the system to act appropriately and in due time. Most challenging will be the highly adversarial multiagent environments of modern battlefields, where the system will likely have to engage with cooperative, agnostic and competitive agents simultaneously (allied agents, civilians and third parties, and enemy agents respectively).

Due to the above reasons, the shift to more complex environments has major ramifications for the types of models that can be employed. Discussions in the Mobility task area demonstrate this well. Automation in Mobility is perhaps one of the most well-explored domains of AI integration (Boulanin, 2016), leading to the development of many autonomous aerial and underwater vehicles. Development on land has, however, been much more challenging: recent experiences with autonomous cars suggest as such (Lin, 2013; Millar, 2014; Boudette, 2017). This contrast has primarily been attributed to the significantly more 'cluttered' nature of land navigation (Canada, 2016; Cummings, 2018). Air and maritime environments are often characterised as very simple, with few if any navigational obstacles. In contrast, land navigation has an abundance of obstacles and specific physics to overcome, such as irregular terrain, vegetation and unstable surfaces, as well as comparatively more agents (both human and automated) to take into consideration.

Land environments for tasks which autonomous weapon capabilities are expected to fulfil will likely be more complex still. Communications-unavailable or -denied areas will invariably not refer to locations held by friendly forces or those which have already been extensively mapped. In practice, these locations could be sprawling urban environments controlled by adversaries. Navigation in such circumstances will be even more challenging than in civil transportation. Matsumura et al. (2014) found that the development of Mobility functions in commercial sectors has the advantage of some level of inherent 'structure' and control in its environment, such as road markings, signs, and right-of-way rules. In contrast, AI in weapon capabilities will have to be able to adapt to very anarchic circumstances, depending on the exact environment in which it is deployed (Sulzbachner et al., 2015).

Sensibly, the Defense Innovation Board (2019, at 66) remarked that a belligerent "does not have control over all aspects of its operational environments". It is due to these challenges that autonomous capabilities currently in operation are generally only deployed in 'controlled' environments. The Israeli Guardium Unmanned Ground Vehicle (Crootof, 2015a; Geiss, 2015), for example, is only used autonomously at the Israel-Gaza border, a location that is well-mapped and relatively static.

These challenges are transposable to the Force domain as well. Accurate targeting and engagement require the agent to be able to properly perceive and analyse its task environment. As is discussed below in Section 5, IHL requires a perpetual distinction be maintained between lawful and unlawful targets. This task, however, is not simple even for humans, as it often requires an analysis of behaviour and context. In this light, perception will likely be the greatest obstacle for automation in complex environments. Boulanin (2016, at 23) confirms that "[i]t is the lack of perceptual intelligence that is impeding the advance of autonomy in some of the most critical applications areas of autonomy in weapon systems".

The problems which were demonstrated within the land navigation domain are indicative of the challenges all systems which are required to engage with complex environments will face. While flight navigation above a dense city will not be necessarily more challenging than normal for an armed aerial vehicle, characterising and analysing ground-based targets will require significantly more sophisticated reasoning. Canada (2016) surmised that moving forward, the utility of any system will remain closely tied to their ability to function in "more complex environments or missions".

4. OPAQUE SOLUTIONS

4.1. Machine learning as a solution

Machine learning models have managed great strides forward due to the recent availability of big data sets (Cukier, 2018). Currently, machine learning sits "at the core of many recent advances in science and technology" (Ribeiro et al., 2016, at 3). Hand-crafted AI generally requires a high degree of problem abstraction and modelling by programmers (Munakata, 2008); experience has shown that this is challenging, sometimes insurmountably so, for tasks and environments too complex or unstructured to map (Abaimov and Martellini, 2020). This is very apparent in the perception domain, where it is often not apparent how to manually represent all the properties of dynamic real-world environments (Wooldridge, 2001). Boulanin (2016, at 23) remarks that "[f]or a number of experts, the solution to designing machines capable of advanced situational understanding lies in the current progress of machine learning". It is highly likely that recourse to the challenges related to environmental and task complexity discussed in Section 3 will be found in the form of resort to machine learning, possibly in conjunction with more symbolic models.

Russell and Norvig (2010) outline three reasons when resort to machine learning models is likely. First, when it is impossible to anticipate all scenarios the agent may be faced with. This is inevitably the case if the capability is tasked with engaging with a dynamic environment. The Dutch Advisory Council on International Affairs (Adviesraad Internationale Vraagstukken, 2015) surmised that capabilities which are expected to classify and engage targets which are not preselected would require machine learning. Second, when it is impossible to anticipate all changes. Russell and Norvig add that this is particularly the case in adversarial scenarios – which most battlefields likely will contain – to which the AI must dynamically adapt for it to be able to 'keep up'. Third, when programming the task manually is too difficult.

This third category includes cases where humans intuitively know what to do but cannot explicitly express how they arrived at this result. Bathaee (2018) draws an apt analogy with how it is impossible to 'explain' how one rides a bike – rather, the common solution is simply to ask the trainee to continue attempting the task until they develop an intuitive understanding. So too would it be easier to 'coach' instead of program such intuition into a machine. As is explored in detail in Section 5, intuition is salient during warfare. Whether 'too many' civilian casualties will be caused by an attack on a military compound or whether the aggressively approaching civilian is actually a threat: human soldiers often must derive their conclusion by intuitively assessing the situation. This matter is often raised as a counterpoint during the political debate by sceptics of autonomous technology on the battlefield (Human Rights Watch, 2012; Sharkey, 2014; Sparrow, 2016), asserting that such decisions can only be made by humans who possess this intuition. Should designers want to implement automation for such tasks, it is therefore likely that machine learning will prove to be the solution.

4.2. Machine learning as an obfuscator

While machine learning has its particular advantages, it also brings specific disadvantages that are relevant to consider. One concerns data reliance, which the Defense Innovation Board (2019) has identified as constituting a significant vulnerability of machine learning-powered agents. Authors have also pointed out dangers of possible bias in the training dataset (Cummings, 2018; Bathaee, 2018; ICRC, 2019) and the inherent vulnerability against data poisoning by adversaries (Wilson, 2020). Abaimov and Martellini (2020) raised the potential problem of data availability. As machine learning models require large quantities of data to construct a robust model (Molnar, 2019), it could be a significant challenge to collect sufficiently sizeable and reliable high-quality datasets for the tasks which the algorithm is expected to perform.

The most significant compromise to make within the context of this paper undoubtedly concerns transparency. Rule-based algorithms have one major virtue: because its designers deliberately set the rules, it is clear why the algorithm arrives at the conclusion that it does (Deng, 2015). The proliferation of machine learning systems has created general concerns about how the opacity generated by this emerging 'black box society' impacts both public and private sectors (Mueller et al., 2019). Biran and Cotton (2017, at §4) write that "contemporary models are more complex and less interpretable than ever; they are used for a wider array of tasks, and are more pervasive in everyday life than in the past; and they are increasingly allowed to make (and take) more autonomous decisions (and actions)".

Molnar (2019, at 13) defines a black box as "a system that does not reveal its internal mechanisms", while Etzioni and Etzioni (2016, at 137) describe them as situations whereby "people are unable to follow the steps these machines are taking to reach whatever conclusions they reach". In essence, they lack interpretability. Outside observers can determine both the input and output variables, but are limited in their understanding of how the system connects them. For convenience, this black box nature is referred to as 'opacity'. Note, however, that opacity exists in certain degrees. Bathaee (2018), for example, subdivides these into strong and weak black boxes, with the former being completely opaque, and the latter allowing a loose ranking of its internal variables and weights. Opacity also can refer to the model or system level. This paper primarily discusses opacity at a system level, i.e., whether the overall operation of a system can be understood (Biran and Cotton, 2017). This distinction is necessary because a system can consist of several models contributing toward an overall decision, and each model also can consist of an ensemble of sub-models. Even if the models can individually be interpreted, the overall system can still be opaque.

Opacity is often the result of modern data-driven algorithms, which obfuscate insights about the data and cause internal relations between nodes to become too intricate for most humans to understand (Mayer-Schönberger and Cukier, 2013). This occurs due to both dimensionality and complexity (Bathaee, 2018; Molnar, 2019). Human visualisation is usually limited to the third dimension and a model operating with 17 variables, leading to a 16-dimensional function, will be impossible to visualise. Complexity is often the result of deep neural networks, which can consist of thousands of neurons, none of which definitively 'determine' the final output, making their processes extremely hard to explain, particularly to non-specialist audiences such as political and military decision-makers (Knight, 2017; Hutson, 2018; Bryce and Parakilas, 2018; Defense Innovation Board, 2019). In addition, one should consider ramifications of systems that are not frozen after training and allowed to continue learning on the field, and if an element of randomisation is implemented. While both of these designs may be effective to allow the system to adapt to enemy strategies and to prevent enemies from exploiting deterministic patterns, they further reduce predictability (Scherer, 2016; Boulanin, 2016; ICRC, 2019). Users and policymakers should be aware of this.

4.3. Compromises

More opaque models are often necessary to improve performance or even make performance possible in the first place for the tasks and environments envisaged for AI in weapon capabilities. If performance is the sole indicator of the system's success, it is inevitable that more and more opaque models will be produced (Molnar, 2019). There are however also important reasons to balance interpretability with performance. This section shortly discusses general incentives found in literature to pursue interpretability. Reasons specifically related to legal requirements are discussed in Section 5.

One main reason to implement interpretability is improving knowledge. Without an explanation of why the system makes a decision, the only thing that is gained is knowledge that the system 'works'. Lombrozo (2006) notes that explanations "are central to our sense of understanding, and the currency in which we exchange beliefs". Studies have indicated that explanations help humans form generalised patterns of inference to understand particular phenomena (Wilkenfeld and Lombrozo, 2015). It is for this reason that interpretability is highly sought-after in the medical field, where it is extremely important to know which variables contribute toward a particular prediction or diagnosis (Knight, 2017). In contrast, it is harder to imagine improving knowledge per se to be an important motivator for transparency in weapon capabilities.

A more relevant incentive for interpretability concerns auditability. Simply put, "you cannot fix what you cannot understand" (Keeley, 2015, at 197). This is extremely important from a military perspective: If an error occurs during the deployment of an automated capability, it is in the belligerent's best interest to determine as quickly and accurately as possible what caused the failure and how to prevent repetition. That is particularly the case if multiple identical systems are currently being fielded with the same vulnerability (Department of Defense, 2015). A related factor is ex ante evaluation. Interpretability during the design phase assists in guarding against possible hidden vulnerabilities, such as algorithmic biases, undesired utility tradeoffs, and the adoption of incomplete proxy objectives (Doshi-Velez and Kim, 2017). These improve the system's reliability, which is also relevant in fulfilling several legal requirements. These are discussed below.

Finally, user trust is frequently cited as an important reason to implement interpretability. Using an autonomous system effectively necessarily entails entrusting important decisions to algorithms (Tolk, 2015). Studies (Ribeiro et al., 2016; Biran and Cotton, 2017) have shown that some degree of explainability is important to achieve social acceptance and that if users do not trust a model or predictions,

they will be wary to use it. Miller (2019) frames explanations in this capacity as fulfilling a persuasive role, i.e., inviting the confidence of human users. User acceptance is desirable from a military perspective as without it, benefits in efficiency which would have been accrued from the automation of a function would be negated due to human users doubting or second-guessing the output (thus re-inserting themselves into the loop and reducing speed and accuracy back to human levels) or refusing to utilise the system at all (nullifying potential military advantages it could have provided). However, this trust should emerge from a genuine and mature understanding of the system's workings and parameters, and not blind faith. According to Mueller et al. (2019, at 5), true persuasion can only come as a "consequence of understanding the how the AI works, the mistakes the system can make, and the safety measures surrounding it". As such, explanations should not only be *persuasive*, but also *accurate*.

5. OPACITY AND IHL

Modern IHL is a specific branch of international law consisting of treaty and customary legal norms which regulate the conduct of hostilities during armed conflict (Gill and Fleck, 2010; Kalindye Byanjira, 2015). These norms include requirements on what characteristics are permissible and impermissible in weapon systems as well as limitations on how such capabilities are used in the field. It is generally accepted that the addition of autonomous functions does not alter State obligations under IHL with respect to that weapon system (ICJ, 1996; Scharre, 2015; Switzerland, 2016).

As a part of international law, IHL imposes the obligation on States to ensure that its requirements are respected in all circumstances (Geneva Convention, 1949, Art. 1; Protocol Additional, 1977, Art. 1(1)). This also entails guaranteeing compliance by its armed forces and that any violations are properly redressed, and if necessary, prosecuted (Geneva Convention, 1949, Art. 49; Fleck, 2013). A significant body of IHL, including the duty to respect and ensure respect, has been recognised as international customary law and therefore binds any State regardless of whether it is party to the main treaties or not (ICJ, 1986; Henckaerts and Doswald-Beck, 2005). As such, conclusions drawn in this section have general applicability and should be taken into consideration by any armed force considering or in the process of integrating opaque AI into its capabilities.

Over the years, discussions have also surfaced with regard to the compatibility of AI with other principles, such as unnecessary harm (Thurnher, 2012; Schmitt, 2013) and the Martens Clause (Mayer, 2015; Geiss, 2015), which will not be addressed here. This section is not meant as an exhaustive exploration of IHL and limits itself to legal requirements which are particularly connected to the problems of opacity presented above. It is explained in each subsection why opacity is potentially problematic for that requirement and its consequences for both designers and users.

5.1. Preparatory concerns: Discrimination and evaluation

Under IHL, a weapon can be intrinsically illegal by virtue of certain inherent qualities which have been deemed undesirable by the international community. A common example of this is poison, which constitutes one of the oldest restrictions in warfare (ICRC, 2006; Boothby, 2016). In modern terms, two main characteristics usually determine if the weapon is inherently unlawful: unnecessary harm and indiscriminateness (Kalshoven, 1990; Parks, 2005). The former has no direct bearing on opaque AI and relates more to the nature of its effectors. An AI-enabled machine designed to fire poison darts would for example be unlawful because of the nature of its projectiles. It would no longer be illegal if the same AI were installed into a machine firing conventional bullets.

Indiscriminateness, on the other hand, is more complicated. The International Court of Justice (ICJ, 1996, at ¶ 78) held in 1996 that it is prohibited to "use weapons that are incapable of distinguishing between civilian and military targets". This requirement is a corollary of the general rule that a distinction must be made at all times between persons and objects of a civilian and military nature (Protocol Additional, 1977, Art. 48; Henckaerts and Doswald-Beck, 2005, Rule 1). Indiscriminate weapons, then, are those that cannot make this distinction, cannot be directed at a specific military objective, or whose effects cannot be controlled (Doswald-Beck, 1997; Schmitt et al., 2006).

As one can see, the main domains related to discrimination are perception and classification. It has been extensively debated both in literature and in political spheres to what extent AI will be able to make these distinctions in the short term (Human Rights Watch, 2012; Thurnher, 2012; Schmitt, 2013; Sassoli, 2014). However, even if we hypothetically assume that the developers have reasonable confidence that the system is capable of correctly classifying objects and agents within its intended working environment and in the fulfilment of its intended task, the nature of this task (perception and classification) makes it very likely that the developers' solution will involve some degree of opacity. This raises possible issues related to preliminary evaluations of the system's reliability.

Switzerland (2016, at ¶8) argues that to determine that a weapon is not indiscriminate, it "must be possible to ensure that its operation will not result in unlawful outcomes with respect to the principle of distinction". This includes both the goal and the means the system selects to achieve that result (Russell and Norvig, 2010; Scherer, 2016). The reliability of capabilities using opaque AI would therefore have to be evaluated before being deployed 'in the wild'. For States Party to the First Additional Protocol to the Geneva Conventions (Protocol Additional, 1977), this duty is attached to the Article 36 requirement that prior to adoption, States should review whether its employment would be prohibited under IHL (which includes the prohibition of indiscriminateness). Although there has been some disagreement (ICRC, 2006; Anderson et al., 2014; Dunlap, 2016) whether the duty of a formal review also extends to non-Parties, the lack of a legal duty to review does not remove the weapon's inherent illegality should it be intrinsically indiscriminate. It is therefore in every State's interest to undertake whatever internal action it deems sufficient to determine the lawfulness of capabilities prior to use (Lawand, 2006). In any case, prior to adoption (either through development or purchase), it is crucial that prospective users are cognisant of the system's reliability – sufficiently so to trust the system on the field make the required distinctions – and that they can provide reasonable guarantees that the algorithm is sufficiently robust in doing so (Keeley, 2015).

Blanchard and Blyler (2016, at 144) define reliability as "the probability that a system or product will perform in a satisfactory manner for a given period of time when used under specified operating conditions". Operators must be able to predict with a high degree of accuracy how the weapon will behave after being deployed. A weapon would not be adequately controllable, and therefore unlawful, if there is more than a remote possibility that it could perform in an unforeseeable way (Doswald-Beck, 1997; Crootof, 2015b). In conventional weapons such as artillery, bombs and missiles, reliability is often associated with the weapon's accuracy, e.g. the rate and frequency of deviations and whether it is very precise or produces a large area-of-effect (McClelland, 2003; ICRC, 2006). During the Yugoslav Wars, for example, the M-87 Orkan was regarded an indiscriminate weapon by the Yugoslav Tribunal in The Hague (ICTY, 2007) because its parameters prevented it from being directed at specific targets in Zagreb. Note that while there has been some disagreement, authors (Crootof, 2015a; Scharre, 2015; Schmitt, 2015) have pointed out that the analysis (i.e., the exact requirements) is tied to the intended working environment. It is therefore important that designers are cognisant and deliberate in what circumstances the capability is intended to function. It is also important to communicate these parameters to users, which can be done in the form of regulations or as part the review process.

Opaque AI complicates the evaluation process because it becomes difficult to make a guarantee that a specific standard of safety has been achieved (Deng, 2015; Abaimov and Martellini, 2020). While reliability estimates can be provided *a priori*, prospective users must also be sufficiently satisfied that the system will maintain the same performance when working with real-world inputs and patterns, which can be very different from the training data (Ribeiro et al., 2016; Cummings, 2018). This is exacerbated when training data was limited to begin with. An exhaustive debug ruling out every single outlying pattern might also be unworkable (Defense Innovation Board, 2019). Doshi-Velez and Kim (2017, at 3) assert that such a guarantee is simply impossible to provide: "For complex tasks, the end-to-end system is almost never completely testable; one cannot create a complete list of scenarios in which the system may fail." It should however be noted that IHL does not require *perfect* accuracy as frankly, a weapon that is 100% accurate all the time has never existed and probably never will (Boothby, 2016). Fenrick (2001) also argued that the assessment of weapon's reliability can only be based on its overall performance, i.e., an isolated weapon malfunction does not render the weapon as a whole illegitimate. The international community should thus seriously consider what (quantitative) standard of reliability would be acceptable with regard to opaque AI when the possibility simply cannot be ruled out that a fringe scenario, possibly a very unexpected one, could cause an undesired output.

Opacity can also complicate the technical development process, particularly in the testing and refining stages. During this phase, the capability's performance is tested, evaluated and fixed. The goal of this process is to identify vulnerabilities both in manufacture and design prior to adoption (Camm, 1993). For software-enabled capabilities, this includes program debugging. Miller (2019) refers to this as the 'examination' purpose of explanations. Just as exams usually require students to explain or motivate their answers, interpretability allows designers to determine whether the algorithm has learned properly or has completely misunderstood the situation. Weaknesses, vulnerabilities and biases can only be debugged when they can be interpreted (Molnar, 2019).

It is highly recommended that interpretability, if necessary, is implemented at this stage of the capability's life-cycle and not after the capability is already deployed (e.g. after an incident occurs). This can be done by developing interpretable models from the outset or incorporating explanations (Defense Innovation Board, 2019; Biran and Cotton, 2017; Miller, 2019). One major reason for favouring this timing is opportunity. Development and testing is the appropriate moment to prevent unreliable items from being passed (Defense Science Board, 1978; Brown, 2010), including those which could be considered indiscriminate as a result. Determining vulnerabilities in this phase can also serve the pre-deployment phase (discussed below), by feeding discovered weaknesses forward to users and enabling them to foresee in which circumstances the capability might no longer fulfil IHL standards. Finally, review processes allow for 'conditional' acceptance, i.e., passing the capability for use but with instructions on when and how to operate them (Daoust et al., 2002; ICRC, 2006). Evaluation results can be incorporated into such instructions to absolutely prevent their use in circumstances found to be incompatible with that system.

5.2. Ex ante concerns: Deployment foreseeability

Even if a weapon is not inherently illegal under IHL, its use can still be unlawful if employed in circumstances that violate certain principles. Three of these are most relevant in the context of this discussion. First, the practical application of the discrimination principle prohibits attacks that are not directed at valid targets (military objectives) (Protocol Additional, 1977, Art. 51(4)(a)). These are distinct from the prohibition on *inherent* indiscriminate weapons discussed above as they can be committed with fundamentally discriminate capabilities (Schmitt, 2006; Schmitt et al., 2006), e.g. use

of precision-guided munitions (PGMs) against an entire civilian neighbourhood because one apartment is being used as a weapon storage. Therefore, even if an AI-enabled capability is deemed not inherently indiscriminate, it must still be employed lawfully by its user. Second, the principle of proportionality, which prohibits attacks that cause an excessive amount of non-military harm (Protocol Additional, 1977, Art. 57(2)(a)(iii)). It is in fact not prohibited under IHL to cause civilian deaths or damage as long as the harm was incidental or a justifiable by-product of a legitimate attack (Quéguiner, 2006). This collateral harm can however not be excessive in relation to the military advantage gained from this attack. As such, to determine proportionality, an accurate prediction is required of the type and extent of damage the attack will inflict. Finally, the principle of precautions requires that prior to an attack, all feasible attempts are made to ensure that distinction is respected and civilian harm is actually kept to a minimum (Protocol Additional, 1977, Art. 57(2); ICTY, 1995).

The application of all these requirements relies heavily on a notion of expectancy and foreseeability. This is indicated by qualifiers such as "may be expected to cause" and "anticipated" (Protocol Additional, 1977, Art. 51(5)(b), 57(2)(a)(iii)). UN Rapporteur Christof Heyns (2013, at ¶70), for example, describes proportionality as requiring "that the expected harm to civilians be measured, prior to the attack, against the anticipated military advantage to be gained from the operation". IHL does not penalise parties who have acted rationally in good faith but nevertheless fell short due to accidents, bad luck, or mistaken (but reasonable) expectations (Fenrick, 2001; Wall, 2002; Schmitt et al., 2006). It does, however, require them take into consideration all information that should be accounted for, including all available intelligence, choice of capabilities and the characteristics of each, environmental properties that can affect the capability's reliability and accuracy, and collateral damage estimations (Rogers, 2000; United Kingdom, 2004).

Several of these factors can prove problematic if a capability with some degree of opaque AI is used. Even if commanders have been primed on the general purpose and functioning of the capability, they might struggle visualising how it will react to specific operational circumstances. 'Environmental properties' usually refers to intuitive factors – e.g. wind, visibility, distance (Schmitt, 2005) – but the analysis could be more complicated with respect to AI. Commanders will have to consider all environmental properties discussed in Section 3.2 and have an understanding of how slight variations in those parameters can affect the weapon's performance. For example, even if the weapon is designed to operate in a multiagent environment, the commander has to consider how the algorithm might perform differently based on the number of other agents present and their orientation (competitive, cooperative, agnostic). Very important also is to consider the environment's adversarial nature (if applicable). What countermeasures by the enemy are foreseeable? How reliable is the system in light of those expected countermeasures?

Unless the commander deploying the capabilities has advanced technical knowledge or is accompanied by an expert capable of performing a trace or audit on the spot, it will not be possible to understand how the algorithm exactly operates. It is also not realistic to expect comprehensive analyses of capabilities at the point when lives are on the line (Blake and Imburgia, 2010), or that meaningful alterations to the algorithm can be performed in the field (Sleesman and Huntley, 2019). For these reasons, it would probably not be reasonable or efficient to push for interpretation at this stage. What the commander needs at this moment in time is a reasonable understanding of how the capability will work in that situation (Dunlap, 2016), which would enable him to make informed estimations of whether the use will result in indiscriminate or disproportionate attacks and any environmental or adversarial factors that could contribute to this.

These considerations increase the value of focusing primarily on interpretability at the previous phase. Keeley (2015) speculates that interpretability at the input-output level would probably be sufficient to

make informed decisions on how percepts from both the task environment and the commander's own inputs (e.g. goals, mission-specific data) influence behaviour. This should be achieved beforehand, during design and adoption. It is also here that feed-forward of conclusions drawn from previous evaluations can be useful, as such conclusions can make commanders aware of specific vulnerabilities. The downing of two friendly aircraft by a Patriot system during Operation Iraqi Freedom is an example of the consequences of failing to feed forward. After the incident, it was found that one of the main contributors to the failure was improper attention to discovering system weaknesses during the design phase (Hawley, 2007, 2011). These fallibilities thus also became unknown to the operators, who could have otherwise acted to prevent the incident. Finally, if the capability is conditionally accepted into use after review, instructions or restrictions can be issued to commanders which succinctly outline conditions wherein the capability may not be used. Such guidelines streamline decision-making at the deployment level through general, concrete rules that commanders simply have to follow without going into too much detail.

5.3. Ex post concerns: Audit and responsibility

Part of the duty under IHL to respect and ensure respect is the obligation to take effective measures against any violations that might occur during operations (Geneva Convention, 1949, Art. 49; Protocol Additional, 1977, Art. 85(1), 86(1)). Taking these measures requires ex post evaluations, which are difficult if the model is opaque, and particularly so if the system itself is destroyed upon impact with the target (e.g. Scharre, 2014; Egozi, 2016; BAE Systems, 2020). In this respect, IHL distinguishes between 'suppression' and 'repression', both of which are relevant to this discussion.

Suppression is wide-ranging and covers "everything a State can do to prevent the commission, or the repetition, of acts contrary" to IHL (Pictet, 1952, at 367). This includes reporting and investigating the breach, taking disciplinary or judicial action, and adapting operational regulations or instructions (Henckaerts and Niebergall-Lackner, 2016). In this respect, interpretable AI is necessary for auditability, i.e., fulfilling the need to know why the machine did what it did after the fact. Keeley (2015, at 214) defines auditability as "the inspection or examination of an entity in order to evaluate or improve its safety or efficiency". Because no weapon is 100% reliable, malfunctions can happen. In such a situation it is the duty of the user to determine the cause of the error and correct it to prevent repetition (or withdraw the weapon if the cause is beyond their capability or authority to fix). For example, knowing the cause of a failure to be a particular adversarial action by the enemy would allow this vulnerability to be removed. If the same algorithm is used in multiple capabilities, the redress should involve all affected systems. Alternatively, the likelihood of similar enemy activity can be taken into consideration by other commanders planning operations with that capability in the future.

Suppression can also include disciplinary action for individuals responsible for breaches or referring them to the proper authorities (Protocol Additional, 1977, Art. 87(3)). For so-called 'grave breaches' (particularly serious infractions of IHL), these individuals must instead be tried penally. It is this latter obligation which constitutes 'repression' (Protocol Additional, 1977, Art. 85(1)). Both suppression and repression relate to a very common restriction of opaque AI: the assignment of blame. When something goes wrong, interpretability helps explain why it did and who should ultimately be held responsible for such failure (Bryce and Parakilas, 2018; Miller, 2019). Because the decision-making process is obscured with opaque AI, such systems "offer us no accountability, traceability, or confidence" (Mayer-Schönberger and Cukier, 2013, at 179). This problem has been ubiquitous in both private and public spheres, ranging from AI used in medical diagnoses, buying, selling and trading, and transportation (Knight, 2017).

The crux of the matter once again is foreseeability. If the internal workings of the algorithm are unknown, its user can claim that he did not (or could not) have anticipated the undesired or unlawful results. In legal terms, this is problematic because the law is not in the habit of punishing persons for outcomes which they did not have knowledge of (Cassese and Gaeta, 2013; Cryer et al., 2014). The International Criminal Court (ICC), for example, places the threshold quite high and requires that the defendant knew that the violation "will occur in the ordinary course of events" (Rome Statute, 1998, Art. 30(3)). However, even less stringent standards such as *dolus eventualis*, where the perpetrator needs to merely be aware of a risk and reconciling himself with it (ICC, 2007, ¶352), will fail if the user operates in complete obscurity. Strict liability – where no degree of fault needs to be proven – is usually limited to product liability and is applied very rarely outside of tort law, and authors (Beard, 2014; Yadav, 2016; Kowert, 2017; Bathaee, 2018) have denounced the option of strict liability in a military context for legal, philosophical and economic reasons. Somewhat eccentric views from certain writers (Sparrow, 2007; Padhy and Padhy, 2019) who have considered transferring accountability to the machine due to its apparent ability to 'think' independently can also be dismissed: the legal provision specifically refers to "persons" (Geneva Convention, 1949, Art. 49) and it is generally agreed that IHL strictly imposes obligations on humans, not their tools (Sassoli, 2014; Sleesman and Huntley, 2019; Defense Innovation Board, 2019).

Two specific elements of legal accountability are particularly problematic vis-à-vis the use of opaque AI. The first concerns intent. Intention in legal terms is usually divided into different stages of deliberation (from recklessness to highly purposive premeditation) which determine if accountability is possible at all and, if yes, the level of responsibility attributed. In criminal law, for example, deliberate offenses are punished more severely than unintended ones. Because machines have no intent, it has been the practice instead to look at the intent of its designers or users. Establishing intent is, of course, straightforward if the program is easily interpretable and was designed to achieve that illegitimate outcome (Schmitt, 2015), in which case users had intended or willingly accepted that end state to occur. In the *US v. Coscia* case (US 7th Circuit, 2017), it was held that the defendant possessed sufficient intent with respect to a high-frequency trade algorithm he had employed to engage in spoofing (market manipulation by artificially placing false bids or orders). However, to establish intent, the Court relied heavily on testimony that the defendant had *requested* the algorithms be programmed with this function in mind. In his commentary, Bathaee (2018) is very sceptical a conviction could have been reached without this proof of direct intent. For example, the algorithm alternatively could have simply been programmed to achieve a goal state of maximising profit, after which it had 'learned' independently that spoofing was an effective way to achieve this end state. If users (and even designers) are ignorant of the program's workings, making the unintended effect completely unforeseeable, there cannot be a high level of intent, if any. Such situations would be particularly problematic with respect to grave breaches, which entail an obligation to repress. In a way, high opacity would thus serve as an undesirable way to shield users and designers by maintaining their ignorance.

The second matter concerns causality. Strictly speaking, the most direct cause of a malfunction would be the percepts which lead to the breach occurring. This has, however, never legally prevented the accountability of persons deemed to have 'caused' those circumstances to take place to begin with. For example, even if a gust of wind is the 'direct' cause of a missile hitting an apartment instead of the adjacent military radio station, the person who chose to launch the missile will still be held as having 'caused' the incident. Under most legal systems, however, this only applies if "the result of the conduct was one that could have been foreseen by a reasonable person" (Bathaee, 2018, at 923). It is questionable how courts will assess the causal chain with regard to situations where the factual direct cause is unknown due to opacity and the only 'known' is the undesired result. Scherer (2016) speculates that in such situations, courts might hesitate in assigning blame to end users.

Because battlefields are adversarial multiagent environments, the cause could have been a user input(s), environmental (which either could or could not have been foreseen by the user) or adversarial (which also may or may not have been foreseeable). The cause might be a complex combination of these factors, and the system might be "so complicated that even the engineers who designed it may struggle to isolate the reason for any single action" (Knight, 2017). It also complicates the allocation of blame, as the adversary may have unwittingly contributed toward the ultimate failure (e.g. a fatal misclassification of a civilian) while intending to achieve a different result (e.g. masking himself from being classified as a target). While actions of other human agents could be construed as a superseding (intervening) cause, this only shifts the problem; even less so than the user, the adversary could not have possibly foreseen how his actions would cause that error. If the algorithm was interpretable, on the other hand, blame could more easily be placed on the user who was aware (or should have been aware) that a certain type of adversarial behaviour could cause the fatal malfunction (Kowert, 2017). By accepting this risk and still deploying the weapon, he would have 'caused' the result.

6. CONCLUDING REMARKS

The rapid development of AI has placed pressure on modern militaries to continually update and improve the precision and speed of their weapon capabilities to remain competitive (Zenko, 2013; Beard, 2014). Simultaneously, there are strong incentives to increase automation for efficiency, safety and versatility, as well as political and economic reasons. Unfortunately, it is precisely the tasks and environments for which AI is needed the most – complex, dynamic and unstructured – which require resort to data-driven algorithms to perform well. The inherent opacity that comes with these models in turn births new challenges, both operational and legal.

All weapons and their use must conform to requirements set by IHL to be employed lawfully. IHL, however, contains many provisions which implicitly or explicitly entail an assumption of foreseeability and predictability. States are required to extensively test the performance of their proposed capabilities to ensure their reliability and that they can perform the necessary discrimination once deployed, which requires evaluations and guarantees. Users are strongly urged to take all necessary precautions to ensure that IHL principles are upheld in the field, which includes gathering as much information as possible on the targets, the target environment and any factors that might influence the weapon's precision. When the commander does not understand how these factors interact with the weapon, however, only so many effective precautions can be undertaken. Finally, if something goes wrong, the problem must be addressed to prevent repetition and, particularly if the violation is severe, persons must be held accountable. These obligations are made difficult to complete if the cause of the problem is unknown, preventing redress and preventing the output to be traced back to the person deemed most responsible. In addition, the lack of foreseeability will frustrate intent and causation tests required for personal responsibility.

In high-risk circumstances such as warfare, interpretability is eminently necessary (Doshi-Velez and Kim, 2017). In such situations, predictions simply "cannot be acted upon on blind faith" (Ribeiro et al., 2016, at 1). There is no explicit prohibition in IHL on the use of capabilities with opaque AI, and it is unrealistic to expect new treaty-based law on the subject within a short time (either through new conventions or amendments to old ones) given the difficulty of obtaining multilateral, let alone global, consensus. Fortunately, existing IHL instruments have proven themselves robust enough to remain relevant and to adapt to many advancements in warfare through their principle-based formulation, which we have applied in this article to opaque AI. Our examination of the different principles indicates that many of the relevant legal obligations become difficult if not impossible to

fulfil if the weapon is allowed to remain fully opaque. As such, an argument could be made that it is not possible for a State to use such systems while remaining fully in accordance with IHL requirements, which is a legal bottom line which may not be crossed. In addition, violating IHL would constitute a breach of an international obligation (Vienna Convention, 1969, Art. 26), which could give rise to State responsibility under international law.

Because interpretability is relevant throughout the capability's entire life-cycle, starting as early as the development and testing phases, it is recommended that interpretability is considered and implemented from the outset. There are however opportunity costs that come with placing more focus on interpretability and transparency. Explanations will have to be both persuasive and accurate, and versatility may have to be reduced by freezing the algorithm after development or removing randomisation. States will have to accept that trade-offs are inevitable, including in development time, cost, and performance (Molnar, 2019). The last trade-off is perhaps counterintuitive, especially if it is argued that increased performance could quantitatively save more lives. More discourse is necessary to determine the appropriate balance between performance and interpretability from a normative perspective. However, it is clear that a singular focus on performance while sacrificing all interpretability would not be in conformity with IHL.

REFERENCES

Abaimov, S. & Martellini, M. (2020). Artificial intelligence in autonomous weapon systems. In M. Martellini and T. Ralf (Eds.), *21st Century Prometheus Managing CBRN Safety and Security Affected by Cutting-Edge Technologies* (pp. 141–177). Cham: Springer Nature Switzerland AG.

Adviesraad Internationale Vraagstukken & Commissie van Advies inzake Volkenrechtelijke Vraagstukken (2015). Autonome wapensystemen: De noodzaak van betekenisvolle menselijke controle. AIV.

Anderson, K., Reisner, D. & Waxman, M. (2014). Adapting the law of armed conflict to autonomous weapon systems. *International Law Studies*, *90*, 386–411.

Anderson, K. & Waxman, M.C. (2013). Law and Ethics for Autonomous Weapon Systems: Why a Ban Won't Work and How the Laws of War Can. In American University Washington College of Law Research Paper No. 2013-11.

Arkin, R.C. (2013). Lethal autonomous systems and the plight of the non-combatant. *AISB Quarterly*, *137*, 1–9.

Arnold, R. (2015). The legal implications of the use of systems with autonomous capabilities in military operations. In A.P. Williams and P.D. Scharre (Eds.), *Autonomous Systems: Issues for Defence Policymakers* (pp. 83–97). NATO: The Hague.

Asaro, P. (2006). What should we want from a robot ethic? *International Review of Information Ethics*, *6*, 9–16.

BAE Systems (2020). Bofors 155mm BONUS Munition. Retrieved June 25, 2020, from www.baesystems.com/en/product/155-bonus.

Bathaee, Y. (2018). The artificial intelligence black box and the failure of intent and causation. *Harvard Journal of Law and Technology*, *31*, 889–938.

Beard, J.M. (2014). Autonomous weapons and human responsibilities. *Georgetown Journal of International Law*, *45*, 617–681.

Biran, O. & Cotton, C. (2017). Explanation and justification in machine learning: A survey. IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI), 8–13.

Blake, D. & Imburgia, J.S. (2010). Bloodless weapons? The need to conduct legal reviews of certain weapons and the implications of defining them as 'weapons'. *Air Force Law Review*, *66*, 157–204.

Blanchard, B.S. & Blyler, J.E. (2016). *System Engineering Management* (5th ed.). Hoboken: John Wiley & Sons, Inc.

Boothby, W.H. (2016). *Weapons and the Law of Armed Conflict* (2nd ed.). Oxford: Oxford University Press.

Boudette, N.E. (2017). Tesla's Self-Driving System Cleared in Deadly Crash. New York Times. Accessed 14 August 2020. Retrieved from www.nytimes.com/2017/01/19/business/tesla-model-s-autopilot-fatal-crash.html.

Boulanin, V. (2016). Mapping the development of autonomy in weapon systems: A primer on autonomy. Stockholm: Stockholm International Peace Research Institute.

Bowne, A.S. (2019). Innovation acquisition practices in the age of AI. *Army Lawyer*, *1*, 75–78.

Boyd, J.R. (1996). C. Richards and C. Spinney (Eds.), *The Essence of Winning and Losing*, Bluffton.

Brown, B. (2010). *Introduction to Defense Acquisition Management* (10th ed.). Fort Belvoir: Defence Acquisition University.

Brown, I.T. (2018). *A New Conception of War: John Boyd, the U.S. Marines, and Maneuver Warfare*. Quantico: Marine Corps University Press.

Bryce, H. & Parakilas, J. (2018). Conclusions and recommendations. In M.L. Cummings, H.M. Roff, K. Cukier, J. Parakilas and H. Bryce (Eds.), *Artificial Intelligence and International Affairs: Disruption Anticipated* (pp. 43–46). London: Chatham House.

Camm, F. (1993). The Development of the F-100-PW-220 and F-1lO-GE-100 Engines: A Case Study of Risk Assessment and Risk Management. N-3618-AF, RAND Note Prepared for the United States Air Force.

Canada (2016). Canadian food for thought paper: Mapping autonomy. CCW Informal Meeting of Experts on Lethal Autonomous Weapon Systems. Accessed 14 August 2020. Retrieved from www.unog.ch/80256EDD006B8954/(httpAssets)/C3EFCE5F7BA8613BC1257F8500439B9F/$file/2016_LAWS+MX_CountryPaper+Canada+FFTP1.pdf.

Cassese, A. & Gaeta, P. (2013). *Cassese's International Criminal Law* (3rd ed.).

Crootof, R. (2015a). The killer robots are here: Legal and policy implications. *Cardozo Law Review*, *36*, 1837–1915.

Crootof, R. (2015b). The varied law of autonomous weapon systems. In A.P. Williams and P.D. Scharre (Eds.), *Autonomous Systems: Issues for Defence Policymakers* (pp. 98–126). NATO: The Hague.

Cryer, R., Friman, H., Robinson, D. & Wilmshurst, E. (2014). *An Introduction to International Criminal Law and Procedure* (3rd ed.).

Cukier, K. (2018). The economic implications of artificial intelligence. In M.L. Cummings, H.M. Roff, K. Cukier, J. Parakilas and H. Bryce (Eds.), *Artificial Intelligence and International Affairs: Disruption Anticipated* (pp. 29–42). London: Chatham House.

Cummings, M.L. (2018). Artificial intelligence and the future of warfare. In M.L. Cummings, H.M. Roff, K. Cukier, J. Parakilas and H. Bryce (Eds.), *Artificial Intelligence and International Affairs: Disruption Anticipated* (pp. 7–18). London: Chatham House.

Dahm, W.J.A. (2012). Killer Drones Are Science Fiction. WSJ. Accessed 14 August 2020. Retrieved from www.wsj.com/articles/SB10001424052970204883304577221590015475180.

Dailymail. Vladimir Putin warns whoever cracks artificial intelligence will "rule the world" (2017). Dailymail. Accessed 14 August 2020. Retrieved from https://www.dailymail.co.uk/sciencetech/article-4844322/Putin-Leader-artificial-intelligence-rule-world.html.

Daoust, I., Coupland, R. & Ishoey, R. (2002). New wars, new weapons?: The obligation of states to assess the legality of means and methods of warfare. *International Review of the Red Cross*, *84*, 345–362. doi:10.1017/S156077550009773X.

Davison, N. (2017). A legal perspective: Autonomous weapon systems under international humanitarian law. UNODA Occasional Papers No. 30.

Defense Innovation Board (2019). AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense Innovation Board. Accessed 14 August 2020. Retrieved from https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF.

Defense Science Board (1978). Report of the Acquisition Cycle Task Force, 1977 Summer Study. Washington, D.C.

Deng, B. (2015). The robot's dilemma: Working out how to build ethical robots is one of the thorniest challenges in artificial intelligence. *Nature*, *523*, 25–27.

Department of Defense (2012). Memorandum. In Defense Science Board (Ed.), The Role of Autonomy in DoD Systems. Washington D.C.: Department of Defense.

Department of Defense (2015). Technical Assessment: Autonomy, Washington, D.C.

Department of Defense (2018). Summary of the 2018 Department of Defense Artificial Intelligence Strategy: Harnessing AI to Advance Our Security and Prosperity.

Doshi-Velez, F. & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. Accessed 14 August 2020. Retrieved from http://arxiv.org/abs/1702.08608.

Doswald-Beck, L. (1997). International humanitarian law and the advisory opinion of the international court of justice on the threat or use of nuclear weapons. *International Review of the Red Cross*, *316*, 35. doi:10.1017/S0020860400084291.

Dunlap, C.J.J. (2016). Accountability and autonomous weapons: Much ado about nothing. *Temple International & Comparative Law Journal*, *30*(1), 63–76.

Egozi, A. (2016). Loitering Weapon Systems – A Growing Demand. IHLS. Accessed 14 August 2020. Retrieved from i-hls.com/archives/73521.

Etzioni, A. & Etzioni, O. (2016). Keeping AI legal. *Vanderbilt Journal of Entertainment & Technology Law*, *19*(1), 133–146.

Fenrick, W.J. (2001). Targeting and proportionality during the NATO bombing campaign against Yugoslavia. *European Journal of Intenational Law*, *12*(3), 489–502. doi:10.1093/ejil/12.3.489.

Ferguson, J. (2001). Crouching dragon, hidden software: Software in DoD weapon systems. *IEEE Software*, Jul/Aug, 105–107. doi:10.1109/MS.2001.936227.

Fleck, D. (2013). *The Handbook of International Humanitarian Law* (3rd ed.). Oxford: Oxford University Press.

Frank, A.U., Bittner, S. & Raubal, M. (2001). Spatial and cognitive simulation with multi-agent systems. In D.R. Montello (Ed.), *Spatial Information Theory*. Morro Bay: COSIT.

Geiss, R. (2015). The International-Law Dimension of Autonomous Weapons Systems. Accessed 14 August 2020, Retrieved from library.fes.de/pdf-files/id/ipa/11673.pdf.

Geneva Convention Relative to the Protection of Civilian Persons in Time of War (1949).

Gill, T.D. & Fleck, D. (2010). *The Handbook of the International Law of Military Operations*. Oxford: Oxford University Press.

Hawley, J.K. (2007). Looking Back At 20 Years Of MANPRINT On Patriot: Observations And Lessons.

Hawley, J.K. (2011). Not by Widgets Alone: The Human Challenge of Technology-Intensive Military Systems. Armed Forces Journal. Accessed 14 August 2020. Retrieved from armedforcesjournal.com/not-by-widgets-alone.

Henckaerts, J.-M. & Doswald-Beck, L. (2005). *Customary International Humanitarian Law, Volume I – Rules*. Geneva: ICRC.

Henckaerts, J.-M. & Niebergall-Lackner, H. (2016). Commentary on the First Geneva Convention. Accessed 14 August 2020. Retrieved from ihl-databases.icrc.org/ihl/full/GCI-commentary.

Heyns, C. (2013). Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions.

Human Rights Watch (2012). Losing Humanity: The Case against Killer Robots.

Hutson, M. (2018). A turtle — or a rifle? Hackers easily fool AIs into seeing the wrong thing. Retrieved August 8, 2020, from ScienceMag website: www.sciencemag.org/news/2018/07/turtle-or-rifle-hackers-easily-fool-ais-seeing-wrong-thing.

ICC (2007). Prosecutor v. Thomas Lubanga Dyilo, ICC-01/04-01/06, Decision (Confirmation of Charges), 29 January 2007.

ICJ. Case Concerning Military and Paramilitary Activities In and Against Nicaragua (Nicaragua v. United States of America) (1986).

ICJ. Legality of the Threat or Use of Nuclear Weapons (1996).

ICRC (2006). *A Guide to the Legal Review of New Weapons, Means and Methods of Warfare: Measures to Implement Article 36 of Additional Protocol I of 1977*. Geneva: ICRC.

ICRC (2019). Artificial intelligence and machine learning in armed conflict: A human-centred approach. Geneva.

ICTY (1995). Prosecutor v Tadić, Decision on the Defence Motion for Interlocutory Appeal on Jurisdiction, 2 October 1995.

ICTY (2007). Prosecutor v Martić, Trial Judgement (IT-95-11), 12 Jun 2007.

Kalindye Byanjira, D. (2015). *Droit International Humanitaire*. Paris: L'Harmattan.

Kalshoven, F. (1990). The conventional weapons convention: Underlying legal principles. *International Review of the Red Cross*, *30*, 510–520. doi:10.1017/S0020860400200065.

Kalshoven, F. & Zegveld, L. (2001). *Constraints on the Waging of War* (3rd ed.). Geneva: ICRC.

Keeley, T. (2015). Auditable policies for autonomous systems (decisional forensics). In A.P. Williams and P.D. Scharre (Eds.), *Autonomous Systems: Issues for Defence Policymakers* (pp. 196–225). NATO: The Hague.

Knight, W. (2017). The Dark Secret at the Heart of AI. Retrieved July 2, 2020, from Technology Review website: www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai.

Kowert, W. (2017). The foreseeability of human-artificial intelligence interactions. *Texas Law Review*, *96*, 181–204.

Lawand, K. (2006). Reviewing the legality of new weapons, means and methods of warfare. *International Review of the Red Cross*, *88*(864), 925–930. doi:10.1017/S1816383107000884.

Lin, P. (2013). The ethics of autonomous cars. The Atlantic. Accessed 14 August 2020. Retrieved from www.theatlantic.com/technology/archive/2013/10/theethics-of-autonomous-cars/280360.

Lin, P., Bekey, G. & Abney, K. (2009). Robots in war: Issues of risk and ethics. In R. Capurro and M. Nagenborg (Eds.), *Ethics and Robotics* (pp. 49–67). Ann Arbor: AKA.

Lombrozo, T. (2006). The structure and function of explanations. *TRENDS in Cognitive Sciences*, *10*(10), 464–470. doi:10.1016/j.tics.2006.08.004.

Marra, W. & McNeil, S. (2013). Understanding "the loop": Regulating the next generation of war machines. *Harvard Journal of Law and Public Policy*, *36*(3), 1–62.

Matsumura, J., Steeb, R., Lewis, M.W., Connor, K., Vail, T., Boyer, M.E., Steinberg, P.S. et al.(2014). *Assessing the Impact of Autonomous Robotic Systems on the Army's Force Structure*. Santa Monica: RAND Corporation.

Mayer, C. (2015). Developing autonomous systems in an ethical manner. In A.P. Williams and P.D. Scharre (Eds.), *Autonomous Systems: Issues for Defence Policymakers* (pp. 65–82). NATO: The Hague.

Mayer-Schönberger, V. & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston: Houghton Mifflin Harcourt.

McClelland, J. (2003). The review of weapons in accordance with article 36 of additional protocol I. *International Review of the Red Cross*, *850*, 397–415. doi:10.1017/S1560775500115226.

Millar, J. (2014). An ethical dilemma: When robot cars must kill, who should pick the victim? Robohub. Accessed 14 August 2020. Retrieved from robohub.org/an-ethical-dilemma-when-robot-cars-must-kill-who-should-pick-the-victim.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. doi:10.1016/j.artint.2018.07.007.

Molnar, C. (2019). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Lean Publishing.

Moya, L.J. & Tolk, A. (2007). *Towards a Taxonomy of Agents and Multi-Agent Systems. 2007 Spring Simulation Multi-Conference*. San Diego: Society for Computer Simulation International.

Mueller, S.T., Hoffman, R.R., Clancey, W., Emrey, A. & Klein, G. (2019). *Explanation in Human-AI Systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI. DARPA XAI Program*.

Munakata, T. (2008). *Fundamentals of the New Artificial Intelligence*. London: Springer.

Osinga, F. (2005). *Science, Strategy and War: The Strategic Theory of John Boyd*. University of Leiden.

Padhy, A.K. & Padhy, A.K. (2019). Criminal liability of the artificial intelligence entities. *Nirma University Law Journal*, *8*, 15–20.

Parakilas, J. & Bryce, H. (2018). Introduction: Artificial intelligence and international politics. In M.L. Cummings, H.M. Roff, K. Cukier, J. Parakilas and H. Bryce (Eds.), *Artificial Intelligence and International Affairs: Disruption Anticipated* (pp. 1–6). London: Chatham House.

Parasuraman, R., Sheridan, T.B. & Wickens, C. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics*, *30*(3), 286–297. doi:10.1109/3468.844354.

Parks, W.H. (2005). Conventional weapons and weapons reviews. *Yearbook of International Humanitarian Law*, *8*, 55–142. doi:10.1017/S1389135905000553.

Pictet, J. (1952). Commentary, I Geneva Convention. Accessed 14 August 2020. Retrieved from ihl-databases.icrc.org/ihl/full/GCI-commentary.

Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (1977).

Quéguiner, J.-F. (2006). Precautions under the law governing the conduct of hostilities. *International Review of the Red Cross*, *88*(864), 793–821. doi:10.1017/S1816383107000872.

Ribeiro, M.T., Singh, S. & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). doi:10.1145/2939672.2939778.

Rogers, A.P.V. (2000). Zero-casuality warfare. *International Review of the Red Cross*, *837*, 165–181.

Rome Statute of the International Criminal Court, 2187 UNTS 90 (1998).

Rosenberg, M. & Markoff, J. (2016). At Heart of U.S. Strategy, Weapons That Can Think. New York Times, A1.

Russell, S.J. & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach* (3rd ed.). New Jersey: Pearson.

Saariluoma, P. (2015). Four challenges in structuring human-autonomous systems interaction design processes. In A.P. Williams and P.D. Scharre (Eds.), *Autonomous Systems: Issues for Defence Policymakers* (pp. 226–248). NATO: The Hague.

Sandoz, Y., Swinarski, C. & Zimmerman, B. (1987). Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949. Martinus Nijhoff.

Sassoli, M. (2014). Autonomous weapons and international humanitarian law: Advantages, open technical questions and legal issues to be clarified. *International Law Studies*, *90*, 308–340.

Scharre, P.D. (2014). Autonomy, "Killer Robots," and Human Control in the Use of Force. Just Security. Accessed 14 August 2020. Retrieved from stsecurity.org/12708/autonomy-killer-robots-human-control-force-part.

Scharre, P.D. (2015). The opportunity and challenge of autonomous systems. In A.P. Williams and P.D. Scharre (Eds.), *Autonomous Systems: Issues for Defence Policymakers* (pp. 3–26). NATO: The Hague.

Scharre, P.D. (2016). Centaur warfighting: The false choice of humans vs. *Automation. Temple International & Comparative Law Journal*, *30*(1), 151–166.

Scherer, M.U. (2016). Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harvard Journal of Law and Technology*, *29*(2), 353–400.

Schmitt, M.N. (2005). Precision attack and international humanitarian law. *International Review of the Red Cross*, *87*(859), 445–466. doi:10.1017/S1816383100184334.

Schmitt, M.N. (2006). War technology and the law of armed conflict. In A.M. Helm (Ed.), *The Law of War in the 21st Century: Weaponry and the Use of Force*. International Law Studies (Vol. 82). Rhode Island: Naval War College Newport.

Schmitt, M.N. (2013). Autonomous Weapon Systems and International Humanitarian Law: A Reply to the Critics. *Harvard National Security Journal*, 1–37.

Schmitt, M.N. (2015). Regulating Autonomous Weapons Might be Smarter Than Banning Them. Retrieved November 5, 2017, from Just Security. Website: www.justsecurity.org/25333/regulating-autonomous-weapons-smarter-banning.

Schmitt, M.N., Garraway, C.H.B. & Dinstein, Y. (2006). *The Manual on the Law of Non-international Armed Conflict, with Commentary*. San Remo: International Institute of Humanitarian Law.

Sharkey, N.E. (2014). Towards a principle for the human supervisory control of robot weapons. *Politica & Società*, 305.

Singer, P. (2009). *Wired for War: The Robotics Revolution and Conflict in the 21st Century*. London: Penguin Press.

Sleesman, R.J. & Huntley, T.C. (2019). Lethal autonomous weapon systems: An overview. *Army Lawyer*, *1*, 32–35.

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy*, *24*(1), 62–77. doi:10.1111/j.1468-5930.2007.00346.x.

Sparrow, R. (2016). Robots and respect: Assessing the case against autonomous weapon systems. *Ethics & International Affairs*, *30*(1), 93–116. doi:10.1017/S0892679415000647.

Statler, I.C. (1993). *Combat Automation for Airborne Weapon Systems: Man/Machine Interface Trends and Technologies. Agard Conference Proceedings 520, Copies of Papers Presented at the Joint Flight Mechanics Panel and Guidance and Control Panel Symposium*. Edinburgh.

Sulzbachner, C., Zinner, C. & Kadiofsky, T. (2015). Passive optical sensor systems for navigation in an unstructured, non-cooperating environment. In A.P. Williams and P.D. Scharre (Eds.), *Autonomous Systems: Issues for Defence Policymakers* (pp. 249–262). NATO: The Hague.

Switzerland (2016). Towards a "compliance-based" approach to LAWS. CCW Informal Meeting of Experts on Lethal Autonomous Weapon Systems, Geneva, 11–15 Apr. 2016. Geneva.

Theunissen, E. & Suarez, B. (2015). Choosing the level of autonomy: Options and constraints. In A.P. Williams and P.D. Scharre (Eds.), *Autonomous Systems: Issues for Defence Policymakers* (pp. 169–195). NATO: The Hague.

Thurnher, J.S. (2012). No one at the controls: Legal implications of fully autonomous targeting. *Joint Force Quarterly*, *67*, 77–84.

Tolk, A. (2015). Merging two worlds: Agent-based simulation methods for autonomous systems. In A.P. Williams and P.D. Scharre (Eds.), *Autonomous Systems: Issues for Defence Policymakers* (pp. 291–317). NATO: The Hague.

UK Ministry of Defence (2011). The UK Approach to Unmanned Aircraft Systems: Joint Doctrine Note 2/11.

United Kingdom. The Joint Service Manual of the Law of Armed Conflict Joint Service Publication 383. (2004).

United States Navy (2019). MK 15 – Phalanx Close-In Weapons System (CIWS). Retrieved December 21, 2019, from www.navy.mil/navydata/fact_display.asp?cid=2100&tid=487&ct=2.

United States Office of General Counsel of the Department of Defense (2015). Law of War Manual, Updated December 2016.

US 7th Circuit. United States v. Coscia, 866 F.3d 782 (2017).

Vienna Convention on the Law of Treaties, UNTS 1155 (1969).

Wall, A.E. (2002). *Legal and Ethical Lessons of NATO's Kosovo Campaign*. Newport: Naval War College.

Wareham, M. (2018). Time is running out. *CCW Report*, *6*(1), 1–4.

Wilkenfeld, D.A. & Lombrozo, T. (2015). Inference to the best explanation (IBE) versus explaining for the best inference (EBI). *Science & Education*, *24*, 1059–1077. doi:10.1007/s11191-015-9784-4.

Williams, A.P. (2015). Defining autonomy in systems: Challenges and solutions. In A.P. Williams and P.D. Scharre (Eds.), *Autonomous Systems: Issues for Defence Policymakers* (pp. 27–64). NATO: The Hague.

Wilson, C. (2020). Artificial intelligence and warfare. In M. Martellini and R. Trapp (Eds.), *21st Century Prometheus Managing CBRN Safety and Security Affected by Cutting-Edge Technologies* (pp. 141–177). Cham: Springer Nature Switzerland AG.

Wooldridge, M. (2001). Intelligent agents: The key concepts. In V. Mařík, O. Štěpánková, H. Krautwurmová and M. Luck (Eds.), *Multi-Agent Systems and Applications II (2001) 9th ECCAI-ACAI / EASSS 2001, AEMAS 2001, HoloMAS 2001*, Selected Revised Papers.

Yadav, Y. (2016). The failure of liability in modern markets. *Virginia Law Review*, *102*, 1031–1100.

Yadron, D. & Tynan, D. (2016). Tesla driver dies in first fatal crash while using autopilot mode. The Guardian. Accessed 14 August 2020. Retrieved from www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk.

Zenko, M. (2013). *Reforming U.S. Drone Strike Policies*. New York: Council on Foreign Relations.