KIRIMLI DR. AZIZ BEY COLLECTED COURSES ON INTERNATIONAL HUMANITARIAN LAW – VOL. II

Onur Dur, Mustafa Can Sati, M. Emre Hayyar

KIZILAY
KÜLTÜR SANAT

Editors

Dr. Onur Dur

Onur Dur is an assistant professor of International Law at the Social Sciences University of Ankara. He completed his PhD degree at the University of Basel, with his thesis on "The Prohibition of Terrorism in International Humanitarian Law". He serves as the General Coordinator of the Istanbul Center for International Law (ICIL). Also, Dr. Dur is a co-founder of the Kırımlı Dr. Aziz Bey International Humanitarian Law Competition and Summer School. Onur holds a Master's degree in International Law from the Geneva Graduate Institute, where he was awarded the Wilsdorf Scholarship. He received his LLB degree from the University of Galatasaray. During his bachelor's studies, Onur participated in the Jean Pictet International Humanitarian Law Competition (Francophone edition-2016) and the René Cassin Human Rights Moot Court (2015).

Dr. Mustafa Can Sati

Dr. Mustafa Can Sati earned his PhD in International Law at Maastricht University, where he completed his doctoral thesis on "Military AI Technologies under International Humanitarian Law: Re-evaluation of Fundamental Principles". He serves as the Administrative Director and a member of the Executive Board of the Istanbul Center for International Law (ICIL) and is a co-founder of the Kırımlı Dr. Aziz Bey International Humanitarian Law Competition & Advanced Summer School. He holds an LL.M. in International and European Law from Ghent University. Previously, he worked as a research assistant in the Public International Law Department at Istanbul Kültür University from 2017 to 2019. He was also selected to participate in the 54th session of the International Law Seminar organised by the United Nations Office at Geneva and the International Law Commission.

M. Emre Hayyar

M. Emre Hayyar is a PhD Candidate in public international law at the University of the West of England, Bristol, United Kingdom. He has been a board member at the Istanbul Center for International Law since August 2017. Mr. Hayyar is a co-founder of the Kırımlı Dr. Aziz Bey International Humanitarian Law Competition and Summer School. Emre holds an LL.M. in Public Law from Istanbul Medipol University (Summa Cum Laude, 2022), an LL.M. in International and European Law from Ghent University (Cum Laude, 2021), and an LL.B. from Istanbul University (2018). He worked as a research assistant in public international law at Istanbul Medipol University between February 2019 and January 2023.

Contents

About The Kırımlı Dr. Aziz Bey IHL Competition 6

Chapter I: 'Testing' the Legality of Autonomous Weapons Systems and AI-based Decision Support Systems under International Humanitarian Law/ Daniele Amoroso 11

1. Introduction 11
2. The First Prong of the Test: Distinction 15
3. The Second Prong of the Test: Proportionality 27

Bibliography 48

Chapter II: International Humanitarian Law and the Regulation of Warfare in the Cyber Domain/ Eitan Diamond 53

1. Introduction 53
2. Adapting Old Law to New Cyber Technologies 56
3. Applying IHL to Technologies and Operations Veiled in Secrecy 59
4. Assessing whether Cyber Operations Occurred in the Context of Armed Conflict 61
5. Environmental Instruments 66
6. Conclusion 76

Bibliography 77

Chapter III: Terrorism and International Humanitarian Law/ Onur Dur 79

1. Historic Evolution of the Prohibition of Terrorism in Armed Conflict until the Geneva Conventions of 1949 79
2. IHL Rules Prohibiting Terrorism 85
3. Relationship between Counter-Terrorism Law and International Humanitarian Law . . . 92
4. Conclusion 94

Bibliography 96

Chapter IV: An Overview of the International Law Commission's Principles on the Protection of the Environment in Relation to Armed Conflict/ Nilüfer Oral 99

1. Introduction 99
2. Protection of Environment in Armed Conflict 100
3. Armed Conflict and the Environment at the UN 102
4. International Legal Framework 102
5. Environmental Instruments 106
6. The International Law Commission 110
7. Conclusion 116

Bibliography 118

Chapter I: 'Testing' the Legality of Autonomous Weapons Systems and AI-based Decision Support Systems under International Humanitarian Law

Daniele Amoroso[1]

Abstract

Autonomous weapons systems (AWS) and AI-based decision support systems (DSS) have been the subject of intense debates within the realms of international law and normative ethics. The increasing deployment of these technologies in the conflicts in Ukraine and Gaza underscores the immediacy and relevance of this discussion. Against this backdrop, the present Chapter aims to take a fresh look at the question of the legality of the use of such systems under international humanitarian law (IHL). This analysis will be carried out through the lens of Art 36 of the First Additional Protocol to the Geneva Conventions, mandating the legal review of new means and methods of warfare. Central to this analysis will be whether AWS and DSS, in their intended uses, align with the principles of distinction, proportionality, and precaution under IHL.

1. Introduction

Over the past years, the debate on the legality of autonomous weapons systems and, more generally, of the use of AI for targeting purposes has grown relentlessly, turning out to be a veritable goldmine for scholars of international law and normative ethics, including –to be fair– the present author. Until recently, such doctrinal hype might have been perceived as somewhat excessive or premature, given that these technologies were either very rudimentary or not yet ready for actual battlefield use. The (alleged) increasing use of AI-based decision support systems and

1 Associate Professor of International and European Law at the University of Cagliari.

autonomous weapons systems in the conflicts in Ukraine[2] and Gaza[3] has reaffirmed the topicality of the issue, to the point that one world-renowned expert in the field has warned of the 'Perilous Coming Age of AI Warfare'.[4]

Against this backdrop, the time seems ripe for a fresh look at whether, and under what conditions, resort to these technologies in armed conflicts is consistent with international humanitarian law (IHL). Before delving into this analysis, a few preliminary remarks are in order.

First, it is important to clarify what is meant by autonomous weapons systems (AWS) and AI-based decision support systems (DSS). Without claiming to take a definitive stance on this point, I will limit myself to providing the working definitions of both types of systems. On the one hand, I will follow the definition of AWS set out by the International Committee of the Red Cross (ICRC), whereby AWS are weapons systems 'that can select (i.e. search for or detect, identify, track, select) and attack (i.e. use force against, neutralize, damage or destroy) targets without human intervention'. [5] DSS, on the other hand, are here understood as systems that rely on AI technologies to support the targeting process by recommending potential targets and courses of action to human operators.[6] The crucial difference between the two, therefore, lies in the fact that, in the case of DSS, the final decision rests in the hands of human operators, while in the case of AWS the decision to engage is autonomously taken by the machine. This distinction,

2 David Hambling, 'Ukrainian AI Attack Drones may be Killing without Human Oversight' (*New-Scientist,* 31 October 2023) <https://www.newscientist.com/article/2397389-ukrainian-ai-attack-drones-may-be-killing-without-human-oversight/> accessed 23 April 2024.

3 Yuval Abraham, "Lavender': The AI Machine Directing Israel's Bombing Spree in Gaza' (+972 Magazine, 3 April 2024) <https://www.972mag.com/lavender-ai-israeli-army-gaza/> accessed 23 April 2024.

4 Paul Scharre, 'The Perilous Coming Age of AI Warfare. How to Limit the Threat of Autonomous Weapons' (*Foreign Affairs*, 29 February 2024) <https://www.foreignaffairs.com/ukraine/perilous-coming-age-ai-warfare> accessed 23 April 2024.

5 International Committee of the Red Cross, 'Views of the International Committee of the Red Cross (ICRC) on Autonomous Weapon System' (Convention on Certain Conventional Weapons (CCW) Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS), Geneva, 11–15 April 2016) <https://www.icrc.org/en/document/views-icrc-autonomous-weapon-system> accessed 23 April 2024.

6 Ruben Stewart and Georgia Hinds, 'Algorithms of war: The Use of Artificial Intelligence in Decision Making in Armed Conflict' (*Humanitarian Law & Policy,* 24 October 2023) <https://blogs.icrc.org/law-and-policy/2023/10/24/algorithms-of-war-use-of-artificial-intelligence-decision-making-armed-conflict/> accessed 23 April 2024.

however, may become less clear-cut in practice in all cases where the human operator is not in a position (e.g. because of time constraints) to question the machine's judgement.

Second, the ensuing analysis will be conducted through the lens of the obligation to review the legality of new weapons, means and methods of warfare, as laid down by Article 36 of the First Additional Protocol to the Geneva Conventions (hereinafter First Additional Protocol or AP I),[7] whereby.

> In the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party.

While it is undisputed that AWS should be subject to such a legality test prior to use, this is not necessarily the case for DSS, which are not weapons per se. One might argue, however, that the integration of AI technologies into the targeting process via DSS could be characterized as a new method of combat, thus subject to legal review as well.[8]

Third, and finally, the assessment of the legality of these new technologies of warfare should take into account the 'conceptual challenge' posed by the performance of the critical function of target identification by an artificial agent.[9] In particular, legal reviews cannot be limited to verifying whether the use of AWS or AI-based DSS is prohibited by IHL or whether their technical characteristics allow humans to behave in compliance with IHL, but should rather aim at establishing whether the system itself is capable of acting lawfully when making 'determinations in the targeting cycle which traditionally are being taken care of by a human operator'.[10] In other words, what is at stake

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

8 On this issue, see International Committee of the Red Cross, 'Artificial Intelligence and Machine Learning in Armed Conflict: A Human-Centred Approach' (2020) 102 (913) *International Review of the Red Cross* 463; Klaudia Klonowska, 'Article 36: Review of AI decision-support systems and other emerging technologies of warfare' [2022] 23 (2020) *Yearbook of International Humanitarian Law* 123.

9 Switzerland, 'Towards a "Compliance-Based" Approach to LAWS' (CCW Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS), Geneva, 30 March 2016) para 23.

10 *Ibid.*

is the system's ability to operate in accordance with the norms of IHL governing the targeting process (the so-called 'law of targeting').[11]

On these assumptions, the Chapter will be structured around the basic tenets of the law of targeting: distinction, proportionality, and precaution. Notably, it will subject AWS and DSS to a threefold legality test:

- Are AWS and DSS, in their normal or intended circumstances of use, capable of complying with the applicable distinction obligations?
- Are AWS and DSS, in their normal or intended circumstances of use, capable of making the proportionality assessments required by IHL?
- Do AWS and DSS, in their normal or intended circumstances of use, hinder compliance with the obligation to take every feasible precaution to avoid (or minimize) civilian harm?

Needless to say, the test will not be carried out in relation to a specific system, as it should be, but in an abstract manner, by making a legal assessment of the current state of the art and research trends in the field.

As it will be seen, each question poses a specific challenge. In particular, section 2 will show that fulfilling the distinction obligations involves a number of complex tasks that, in most scenarios, cannot be performed autonomously by AWS and DSS, unless they are provided with the capability to acquire situation awareness. Section 3 will investigate whether and how the thrust of the principle of proportionality, i.e. the obligation to strike a reasonable balance between heterogeneous values (collateral damage and military advantage), can actually be translated into a machine algorithm. Section 4 will show that the positive obligation to take precautionary measures under IHL imposes important constraints on how States should conceive of the distribution of control privileges between humans and machines. Section 5 will wrap up the discussion and make some concluding observations as to whether, to what extent, and under what conditions AWS and DSS can pass the (theoretical) legality test set out in this Chapter.

11 This requirement could be somehow based on Art 51(4)(b) and (c) AP I, which respectively prohibit the use of methods or means of combat which "cannot be directed at a specific military objective" and "cannot be limited as required by th[e] Protocol". On such a conflation between targeting law and weapons law (rectius: means and methods of warfare law). See also Natalia Jevglevskaja, *International Law and Weapons Review. Emerging Military Technology under the Law of Armed Conflict* (CUP 2021) 207–238.

2. The First Prong of the Test: Distinction

A key acquisition of modern international law is that, during armed conflicts, certain persons and objects must be protected from direct attacks. Accordingly, weapons that 'are incapable of distinguishing' between legitimate targets and protected persons/objects should be viewed as 'indiscriminate' and thus forbidden.[12] As far as conventional weapons are concerned, this requirement is understood as prohibiting the use of weapons that do not allow the human user to make such distinctions, on the assumption that a weapon, 'being an inanimate object', cannot perform an activity that implies 'thought'.[13] This could no longer be valid for AWS and DSS, whose defining feature is precisely that of carrying out the critical function identifying targets on their own.[14] Regardless of whether this activity actually requires 'thought', as it has been suggested, the fact remains that the legality of AWS and DSS passes, in the first place, through an assessment of their ability to direct themselves against legitimate targets.

Although its 'seeds' might be traced back to the 1868 Saint Petersburg Declaration,[15] the principle of distinction was expressly formulated only in 1977 with the adoption of the First Additional Protocol to the Geneva Conventions, whose Article 48 (Basic Rule) posits:

> In order to ensure respect for and protection of the civilian population and civilian objects, the Parties to the conflict shall at all times distinguish between the civilian population and combatants and between civilian objects and military objectives and accordingly shall direct their operations only against military objectives.[16]

---

12 *Legality of the Threat or Use of Nuclear Weapons*, Advisory Opinion [1996] ICJ Reports 226, para 78. See also Art 51(4) AP I.

13 Louise Doswald-Beck, 'International Humanitarian Law and the Advisory Opinion of the International Court of Justice on the Legality of the Threat or Use of Nuclear Weapons' [1997] 316 *International Review of the Red Cross* 35, 39.

14 Kjølv Egeland, 'Lethal Autonomous Weapon Systems under International Humanitarian Law' [2016] 85 *Nordic Journal of International Law* 89, 97.

15 Declaration Renouncing the Use, in Time of War, of Explosive Projectiles Under 400 Grammes Weight. Saint Petersburg, 29 November/11 December 1868. See, in particular, the statement whereby the "only legitimate object which States should endeavor to accomplish during war is to weaken the military forces of the enemy" and that for this "purpose it is sufficient to disable the greatest possible number of men".

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

The same rule is reiterated by Articles 51(2) and 52(1) AP I, which dealing with, respectively, the 'protection of civilian population' and the 'general protection of civilian objects'. The picture is further enriched by a number of protective rules which, while not aimed at immunizing the civilian population from direct attacks, raise analogous issues of distinction. The most important of these is the norm prohibiting attacks on persons hors de combat, namely –according to the definition of Art 41(2) AP I– those enemies who clearly expresses an intention to surrender; or are defenceless because of unconsciousness, shipwreck, wounds or sickness; provided that they abstain from any hostile act and do not attempt to escape.

2.1. Unresolved Issues: Direct Participation in Hostilities and 'Dual-Use' Objects

At a first sight, the prescriptive content of the principle of distinction seems crystal-clear: do not attack civilians and civilian objects or, to put it the other way around, do attack only combatants and military objectives.

Contemporary IHL, however, features the persistence of a number of grey areas, especially in relation to asymmetric warfare, where hostilities are generally conducted in urban settings and armed actors are closely intermingled with civilians, who do not always abstain from participating in hostilities. The latter problem is addressed in very general terms by Article 51(3) AP I, which suspends civilians' immunity from attacks 'for such time as they take a direct part in hostilities', but fails to offer proper directions as to what counts as 'direct participation in hostilities'.

In an attempt to shed light on this question, in 2009 the ICRC issued an *Interpretive Guidance on the Notion of Direct Participation in Hostilities under International Humanitarian Law,* authored by the then ICRC's Legal Adviser Nils Melzer.[17] The ICRC identified three cumulative criteria to establish whether an act is one of 'direct participation in hostilities':

- The act must reach a 'threshold of harm', namely it must be likely to adversely affect the military operations or capacity of

17 Nils Melzer, *Interpretive Guidance on the Notion of Direct Participation in Hostilities under International Humanitarian Law* (ICRC 2009).

a belligerent party or, alternatively, to harm persons or objects protected against direct attack, and

- There must be a direct causal link between the act and the harm likely to result either from that act, or from a coordinated military operation of which that act constitutes an integral part, and
- The act must be specifically designed to directly cause the required threshold of harm in support of a party to the conflict and to the detriment of another (belligerent nexus)[18]

The same criteria are employed by the ICRC also to identify members of irregular armed forces, who are those carrying out acts of direct participation in a continuous manner (so-called 'continuous combat function' criterion).[19]

To be clear, these criteria have not yet been universally accepted.[20] In the ensuing analysis, however, the authoritative view of the ICRC will be taken as a reference, with the caveat that, even if one adopted the more liberal approaches proposed by the critics of the Interpretive Guidance, this would not significantly diminish the perceptual and evaluative challenges posed to AWS and DSS by the notion of direct participation in hostilities.

Another tricky issue in IHL is posed by objects sometimes labelled as 'dual-use', namely those civilian objects that, similarly to civilians taking part to hostilities, lose their immunity from attacks because, by their use or purpose (i.e. by their actual or intended use), they make an effective contribution to the military action of the enemy.[21] The construction of the 'purpose' and 'use' criteria is indeed featured by a number of grey areas as well. This is illustrated by an oft-cited example: television and radio broadcasting stations, like the Radio Television of Serbia, whose

18 *Ibid.*

19 Nils Melzer, 'Keeping the Balance between Military Necessity and Humanity: A Response to Four Critiques of The ICRC's Interpretive Guidance on the Notion of Direct Participation in Hostilities' [2010] 42 *New York Journal of International Law & Policy* 831, 854.

20 See, for instance, the critical remarks by Kenneth Watkin, 'Opportunity Lost: Organized Armed Groups and the ICRC "Direct Participation in Hostilities" Interpretive Guidance' [2010] 42 *New York Journal of International Law & Policy* 641; Michael Schmitt, 'Deconstructing direct participation in hostilities: The constitutive elements' [2010] 42 *New York Journal of International Law & Policy* 697; William H Boothby, "'And for Such Time As": the Time Dimension to Direct Participation in Hostilities' [2010] 42 *New York Journal of International Law & Policy* 741.

21 Christine Byron, 'International Humanitarian Law and Bombing Campaigns: Legitimate Military Objectives and Excessive Collateral Damage' [2011] 13 (2010) *Yearbook of International Humanitarian Law* 175, 183–188.

headquarters in Belgrade were bombed in 1999 by NATO forces. While it is unlawful to target them only to halt the enemy's propaganda machine (because that would constitute a too remote military advantage), television and broadcasting stations constitute legitimate targets when they are part of the Command, Control and Communications network of the enemy's army, as it was arguably the case of the Radio Television of Serbia.[22]

2.2. The Problem of Situation Awareness

The foregoing provides us with an adequately representative sample of identification tasks that AWS and DSS should be legally required to perform with a view to complying with distinction obligations under IHL. To describe in a synthetic fashion what it takes to properly perform these tasks in a warfare scenario, it is sometimes used a notion borrowed from ergonomics (but more and more resorted to in military studies): that of situation awareness (SA), i.e. the knowledge of 'the state of a dynamic environment'.[23] The performance of these activities, in other terms, requires not only that the pertinent rules of international law are properly coded into algorithms, but also –and even more importantly– that these systems are endowed with the capability to autonomously 'assess' the surrounding environment.[24]

A helpful model to approach this problem was set out by Mica Endsley, former Chief Scientist of the US Air Force, who dissected it into three interrelated and hierarchically structured components: 'Perception of the elements in the environment' (Level 1), 'Comprehension of the current situation' (Level 2), 'Projection of future status' (Level 3).[25] The usefulness of this model will be apparent if we consider the case of the identification of a surrendering combatant. In order to comply with the

22 This point was clearly made by the Committee Established to Review the NATO Bombing Campaign Against the Federal Republic of Yugoslavia, *Final Report to the Prosecutor by the Committee Established to Review the NATO Bombing Campaign against the Federal Republic of Yugoslavia* (2000) paras 47, 55 and 77 (OTP Report).

23 MR Endsley, 'Toward a Theory of Situation Awareness in Dynamic Systems' (1995) 37(1) *Human Factors* 32, 36.

24 Lucy Suchman, 'Situational Awareness and Adherence to the Principle of Distinction as a Necessary Condition for Lawful Autonomy.' (Briefing paper, (CCW Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS), Geneva, 12 April 2016).

25 Endsley (n 22) 36.

prohibition of attacking persons hors de combat, an AWS (or a DSS) should be able to perform the following sub-tasks (*a*) to recognize a certain object as a human combatant and to discern his/her gestures (perception of the elements of the environment, both in their static and dynamic dimensions); (*b*) to infer from his/her appearance and gesture, as well as from the context, that (s)he is conveying a message of surrender (comprehension of the current situation); (*c*) to predict, on the basis of the combatant's behavior, whether (s)he does or does not intend to re-engage in hostile acts or attempt to escape (projection of a future status).

2.2.1. Perception of Elements in the Environment (Level 1 SA)

A *sine qua non* condition for situation awareness is the perception 'of the status, attributes, and dynamics of relevant elements in the environment'.[26] Battlefield scenarios may consist of a myriad of elements, but of course not all of them need to be perceived by the decision-maker, be it a human or a machine. For our purposes, therefore, it is sufficient that the system is able to recognize the elements whose perception, misperception, or non-perception could affect the lawfulness of a targeting decisions. How demanding this will be for a system depends on several variables. In this respect, one may distinguish between tractable recognition tasks and perceptual challenges that are still unmatched by current AI research.

A distinction task that is typically tractable by existing technologies is the recognition of so-called 'cooperative' objects. Cooperative objects 'are those that are actively emitting a signal, which makes them easier to detect'.[27] A classic example is the radar that allows enemy forces to find its position by emitting electromagnetic pulses, which are discernible by passive sensors (i.e. sensors that only receive information from the surrounding environment). The autonomy of the Israeli Harpy NG anti-radiation munitions is based on this relatively simple technology that, however, does not allow the weapon to establish whether the source of radio emission is located in the proximity of a protected object (e.g. a school).[28] Quite interestingly, there are some norms of IHL

26 *Ibid.*

27 Paul Scharre, *Army of None. Autonomous Weapons and the Future of War* (WW Norton & Company 2018) 84.

28 NE Sharkey, 'The Evitability of Autonomous Robot Warfare' (2012) 94 (886) *International Review of the Red Cross* 787, 788.

that promote 'cooperation' from protected objects in order to prevent them from being targeted. Annex I to AP I, in particular, contains very detailed provisions regulating the emission of specific light, radio, and electronic signals by medical aircrafts, vessels, and vehicles with a view to making them easily detectable by the Parties to the conflict.[29]

Yet, theatres of war are often populated by 'non-cooperative' objects, be them targetable (e.g. a tank) or protected from attacks (e.g. a school bus), which do not send out signals that may be detected by passive sensors. In this case, the system should be able to 'sense' the operational area; to 'discern' the objects of interest from the surrounding environment; and to 'recognize' the objects, by associating each of them to a certain class (e.g. tanks, non-military vehicles, combatants). Some norms of IHL are in fact aimed at assisting this activity, insofar as they require the display of easily discernible distinctive signs in order to flag that a certain object is not a lawful target. One could think, in this respect, to the provisions governing the use of the emblem of the red cross (and the like) for medical service,[30] the blue shield for cultural property,[31] and the three bright orange circles signaling 'works and installations containing dangerous forces'.[32] The same goes for the combatants' obligation 'to distinguish themselves from the civilian population [e.g. by wearing uniforms] while they are engaged in an attack or in a military operation preparatory to an attack'.[33] In these cases, object perception would boil down to the relatively affordable task of identifying the distinctive signs.

The identification of fast-approaching incoming threats, such as missiles, rockets or aircrafts, is not particularly demanding as well, as it is generally carried out in the light of two simple criteria: speed and trajectory. These are detected with an acceptable degree of reliability by air defense systems.[34] More complex, but still achievable, is the autonomous

29 Annex I to AP I: Regulations concerning identification [1977], Arts 6–8, 10–11.

30 ICRC, 'Emblem: Relevant Articles of the 1949 Geneva Conventions and their Additional Protocols' (1 November 2008) <www.icrc.org/eng/resources/documents/misc/emblem-ihl-011108.htm> accessed 23 April 2024.

31 *Hague Convention for the Protection of Cultural Property in the Event of Armed Conflict* 1954, Arts 16 and 17.

32 Art 56(7) AP I.

33 Art 44(3) AP I.

34 Vincent Boulanin and Maaike Verbruggen, *Mapping the Development of Autonomy in Weapon Systems* (SIPRI 2017) 37–39.

qualification of objects as military objectives *by their nature,* such as tanks, navy ships, helicopters.[35] The UK Brimstone fire-and-forget munition, for instance, is reportedly able to sense and discern the objects in a given operational environment through a high-millimetric band seeker, which provides a high-resolution radar return image of them. These images are evaluated by the system's target recognition algorithms, through a comparison with the templates stored in the target identification library, in order to recognize armored vehicles and distinguish them from other fixed or moving objects, such as houses or cars.[36]

Let us now consider the hard cases. The evaluation as to whether a civilian is directly taking part in hostilities presupposes, at the very least, not only that an object is recognized as a human being, but also the discernment of all the tools (s)he is interacting with (e.g. a rifle), the kind, features and addressees of the actions (s)he is performing (e.g. shooting against friendly forces), as well as the contextual elements that could offer an alternative explanation to his/her behavior (e.g. the need to react to unlawful acts of violence). Likewise, an assessment aimed at qualifying an object as a military objective by its use or purpose entails the ability to perceive a plurality of objects and dynamics thereof (e.g. in the case of a school used as a deposit of ammunitions: a building, a group of enemy combatants transporting munitions into it, trucks parked outside the school, and so on). More in general, counterinsurgency operations in urban settings, where the close intermingling between legitimate targets and protected persons/objects exponentially raises the range of elements to be 'perceived' (i.e. civilians, combatants, military and non-military vehicles, weapons, protected objects, and all the dynamics thereof), make it particularly demanding for an artificial agent to autonomously acquire even the lowest level of situation awareness.

That notwithstanding, recent progresses in the field of deep learning are such that one cannot rule out that a system provided with a sophisticated learning architecture could be trained to discern and recognize all the objects in the surrounding environment, by feeding it with hundreds of thousands of footages showing the operational environments and the

35 Alan Backstrom and Ian Henderson, 'New Capabilities in Warfare: An Overview of Contemporary Technological Developments and the Associated Legal and Engineering Issues in Article 36 Weapons Reviews' (2012) 94(886) *International Review of the Red Cross* 483, 496.

36 Boulanin and Verbruggen (n 33), 49–50.

objects it can find within it (insurgents, military vehicles, civilian cars, kids, and so for).[37] This would be, however, only the bottom rung of the 'situation awareness' ladder. Things become even trickier for AWS when we pass to consider the upper levels.

2.2.2. Comprehension of the Current Situation (Level 2 SA) and Projection of Future Status (Level 3 SA)

Visual recognition would only provide the system with a disjointed knowledge of the objects present in the battlefield, which could prove a very poor basis for taking a lawful target decision. Let us consider the following scenario:

Insurgents (without uniforms) are firing against friendly forces a battery of mortars, located in the middle of an urban street. Mortars' munitions are stored in a nearby school and are transported to the fire position by a civilian truck at regular intervals. In the meanwhile, a bus has approached the area in order to evacuate civilian buildings on the street where armed confrontation is taking place.

Even if endowed with technology granting a high level of accuracy in object recognition, a system would simply identify a disaggregated picture composed of friendly forces, a battery of mortars, a moving truck, a number of buildings (including the school), a bus, and people without uniforms scattered around the area of operations. That could be enough to qualify the mortars as military objectives by nature and friendly forces as protected from attack, but it would say nothing as to the legal status of the other objects, thus preventing the system from conducting a proper distinction analysis. To have a more complete picture as to what is a lawful target and what is not, indeed, analytical perception of the environment should be complemented by an activity of synthesis, aimed at comprehending the dynamics of the elements and predicting future statuses thereof. Only this activity would allow the system to characterize the school and the truck as military objectives objects which, by their use, 'make an effective contribution to military action'; the driver of the truck as taking direct part to the hostilities; and the bus and the other buildings as objects protected from direct attacks.

37 This seems to be the main goal pursued by the US DoD Project Maven. See Julia Press, 'Inside Project Maven, the US Military's Mysterious AI Project' (*Bloomberg*, 29 February 2024) < https://www.bloomberg.com/news/articles/2024-02-28/inside-project-maven-the-us-military-s-flagship-ai-project-big-take-podcast?embedded-checkout=true> accessed 23 April 2024.

Enabling AWS to acquire these levels of situation awareness appears to be a challenge far more demanding than that posed by object perception.[38] In this connection, one may observe that human warfighters generally follow two ways to reach full situation awareness: feature-matching and story-generation. Feature-matching consists in comparing the available data with mental models built on the basis of training and past experience. When no match with existing models is found, the decision-maker generates a causal story in order to give meaning to the cues coming from the environment and to forecast future events.[39]

No doubt, feature-matching could be technically implemented into AWS to enhance their situation assessment capabilities. Notably, a number of templates, or 'signatures', could indeed be loaded into the system's long-term memory with a view to guiding the assessment as to whether a civilian is directly participating in hostilities, an object is a legitimate military objective by reason of its use or purpose, or a pattern of behavior conveys an intention of surrender.[40] It is of the utmost importance, however, that these templates accurately reflect IHL and do not allow leeway for dangerous generalizations. The US practice on signature strikes is a stark reminder of the risk that the adoption of loose targeting criteria (e.g. all males of military-age located in the area of enemy's activities) sets in fact the conditions for indiscriminate attacks against the civilian population.[41] This risk has been confirmed in the recent conflict in Gaza, where Israel Defense Forces reportedly relied on an AI-based DSS ('Lavender') to identify low and medium rank members of the military wings of Hamas. According to some sources, Lavender was trained to establish the likelihood that a Gaza resident was a militant based on a number of pretty controversial signatures, 'such as being in a Whatsapp group with a known militant, changing cell phone every few months, and changing addresses frequently'.[42]

38 David Akerson, 'The Illegality of Offensive Lethal Autonomy', in Dan Saxon (ed.), *International Humanitarian Law and the Changing Technology of War* (Martinus Nijhoff Publishers, 2013) 65, 76–82.

39 George L Kaempf, Steve Wolf and Thomas E Miller, 'Decision Making in the AEGIS Combat Information Center' [1993] 37 *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 1107, 1109.

40 The South Korean sentry robot SGR-A1, now retired, was reported to be able to recognize surrender motions (ie arms held high). See Boulanin and Verbruggen (n 33) 45.

41 Kevin Jon Heller, "One Hell of a Killing Machine': Signature Strikes and International Law' [2013] 11(1) *Journal of International Criminal Justice* 89.

42 Abraham (n 2).

In any case, one cannot take it for granted that the entire range of possible dynamics of unstructured warfare environments can be successfully pre-set into templates to be installed into AWS. In cases not covered by the templates, full situation awareness should therefore be achieved through story-generation (or 'abductive reasoning'), namely by piecing together incomplete information and make 'best guess' decisions. Yet, approaching human-like performances in this field is, unlike object recognition, an enduring problem that machine learning techniques has so far failed to address in an adequate manner.[43]

The foregoing suggests that it cannot be expected, at least in the near future, that AWS and DSS will be endowed with the ability to acquire the levels of situation awareness needed to comply with the principle of distinction in complex battlefield scenarios. This is not to mean, however, that this technological development can never be achieved, as it is sometimes suggested. For all we know, lack of situation awareness is a contingent situation that may no longer hold at some future time.

2.3. Is Situation Awareness Really Needed?

In the debate on the legal implications of autonomy in weapons systems, the need to enable AWS to achieve an adequate level of situation awareness has sometimes been downplayed (if not denied altogether). In general, two arguments have been generally put forward. First, it is submitted that issues of distinction are dealt with by human commanders at the planning stage of the targeting process. Second, it is argued that problems of distinction do not necessarily arise in every battlespace.[44]

The *first objection* is specific to AWS, as DSS are designed to always leave the final decision to engage the recommended target to a human operator.[45]

43 Hossein Chegini, 'Unraveling Abductive Reasoning: Chatgpt's Struggles and Limitations Exposed' (Medium, May 2023) <https://medium.com/coding-intelligence-ai-development-and-analysis/unraveling-abductive-reasoning-chatgpts-struggles-and-limitations-exposed-27fd4da6c60b> accessed 23 April 2024.

44 Michael N Schmitt and Jeffrey S Thurnher, "'Out of the Loop": Autonomous Weapon Systems and the Law of Armed Conflict' [2013] 4(2) *Harvard National Security Journal* 231.

45 Another issue is whether the human operator is put in a position to *meaningfully* intervene in the decision-making process, or merely rubber-stamps the recommendation of the DSS (Daniele Amoroso and Guglielmo Tamburrini, 'Toward a Normative Model of Meaningful Human Control over Weapons Systems' [2021] 35(2) *Ethics & International Affairs* 245). The latter scenario is exemplified by the use of the Lavender system by the Israel Defense Forces. It has indeed been reported that human operators limited themselves to checking whether the proposed target was a man or a woman, since women cannot be part of the Hamas military wing. See Abraham (n 2).

It is based on the assumption that it is at the planning stage that the legal assessment is made as to whether a particular object (e.g. a bridge) constitutes a military objective under IHL. From this perspective, AWS will therefore limit themselves to merely executing targeting decisions assumed by humans in the so-called 'wider loop'.[46]

Indeed, the fact that human decision-makers will always (or at least for a very long time) have a prominent role in planning an attack may be relevant in relation to *deliberate* targeting, which prosecutes targets that are 'known to exist in an area of operations and can be mapped to decisive points on line(s) of operation',[47] but it is largely unhelpful with regard to *dynamic* targeting, which prosecutes targets of opportunity such as unplanned targets (i.e. military objectives which were known to exist, but whose location and/or importance is grasped only after the attack was initiated) and unanticipated targets (i.e. military objectives which were unknown to exist at the planning stage).[48]

Moreover, exclusive reliance on the planning phase seems to overlook that the obligation to 'distinguish between the civilian population and combatants and between civilian objects and military objectives' should be complied with 'at all times',[49] not only at the time when an attack is planned. This aspect is reiterated and expounded on by Article 57 AP I, which obliges the warring parties to take 'constant care […] to spare the civilian population, civilians and civilian objects' and, more in detail, to cancel or suspend an attack 'if it becomes apparent that the objective is not a military one or is subject to special protection'. The latter obligation is imposed upon anyone who has effective control over the attack, who may well be the person in charge of executing the targeting decision, especially if communication with the chain of command are severed.[50]

46 See Jeffrey S Thurnher, 'Examining Autonomous Weapon Systems From a Law of Armed Conflict Perspective', in Hitoshi Nasu and Robert McLaughlin (eds.), *New Technologies and the Law of Armed Conflict* (Springer, 2014) 213, 223–225; Mark Roorda, 'NATO's Targeting Process: Ensuring Human Control over and Lawful Use of 'Autonomous' Weapons', in Andrew Williams and Paul Scharre (eds.), *Autonomous systems: Issues for defence policymakers* (NATO, 2015) 152; Merel A C Ekelhof, 'Lifting the Fog of Targeting: "Autonomous Weapons" and Human Control Through the Lens of Military Targeting' [2018] 71(3) *Naval War College Review* 61.

47 Angus G Houston, 'Australian Defence Doctrine Publication 3.14 – Targeting' (2 February 2009) para 1.11.

48 *Ibid*, para 1.12.

49 AP I, Art 48(1).

50 Ian Henderson, *The Contemporary Law of Targeting* (Brill, 2009) 235.

In case of deliberate targeting carried out through AWS, the person executing the targeting decision (and thus the one legally obliged to ascertain whether the locked-on target actually constitutes a military objective) will be replaced by the weapons systems, which raise again the problem of endowing the latter with SA capabilities. This issue is all the more pressing in case of AWS enabled to loiter on the target area for extended periods of time, as the temporal distance between the planning of the attack and its actual execution makes it highly probable that some change may occur on the battlefield that should be factored in distinction analysis.

The *second objection* may also be of some relevance to DSS, since it could be contended that, in scenarios where there are no issues of distinction, the system would be able to recommend *lawful* targeting decisions even without acquiring SA. While it is true that the need to distinguish civilians and civilian objects does not arise on any battlefield, the rule prohibiting attacks against enemies *hors de combat* which as we have seen poses non-trivial perceptual and evaluative challenges, also applies in a warfare scenario where civilians are not present.[51] The only case in which the use of an AWS or a DSS that is incapable of acquiring high levels of SA would not be problematic under IHL would be if the system were programmed to 'hunt' uninhabited enemy objects, such as armed drones, in uncluttered environments. In this case, all that would be required is for the system to reliably discriminate between the assigned targets and any other object.

2.4. Summary of the Discussion

To be used in accordance with IHL distinction obligation, AWS and DSS should be equipped with the capability to acquire and maintain situation awareness. Currently fielded systems are able to reach Level 1 SA (Perception) in terms of recognizing targets that are discernible on the basis of shape, heat and motion patterns in uncluttered environments. Making lawful targeting decisions in more complex scenarios generally requires the capability to autonomously achieve Level 2 and Level 3 SA (Comprehension and Prediction), which remains a major challenge for AI research.

51 Rob Sparrow, 'Twenty Seconds to Comply: Autonomous Weapon Systems and the Recognition of Surrender' [2015] 91 *International Law Studies* 699, 712.

The obligation to provide AWS and DSS with situation assessment capabilities has not necessarily the same breadth for each and every system and use thereof. Notably, higher levels of situation awareness are required if the system is (*a*) deployed in a cluttered environment, i.e. characterized by the intermingling of protected objects and military objectives; or (*b*) tasked to prosecute targets of opportunity (dynamic targeting); or (*c*) allowed to select and/or engage human targets and/or human-inhabited objects. Conversely, low levels of SA may be sufficient for systems that (*a*) are deployed in uncluttered scenarios; (*b*) are programmed to attack anticipated or known targets within the assigned operational environment; and (*c*) are assigned to attack only uninhabited military objects.

3. The Second Prong of the Test: Proportionality

While an early manifestation of the proportionality principle in IHL may be found in Article 24(4) of the (never adopted) 1923 Hague Rules of Air Warfare, its codification as a legally binding norm is due to Article 51(5)(b) AP I, which labels as 'indiscriminate', and thus prohibited, those attacks that are

> expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated.[52]

A provision to the same effect is contained in Article 57(2)(a)(iii) AP I, and both are reinforced by Article 57(2)(b), which obliges warring parties to cancel or suspend an attack 'if it becomes apparent' that it may be expected to cause disproportionate collateral damage. Although the Second Additional Protocol to the Geneva Conventions does not envisage a similar provision in relation to non-international armed conflicts, the ICRC provided convincing evidence that customary law demands respect for the principle regardless of the nature of the conflict.[53]

From a plain reading of the norms embodying the principle, it emerges that proportionality is not concerned with actual outcomes but with

52 AP I, Art 51(5)(b).

53 J-M Henckaerts and Louise Doswald-Beck, *Customary International Humanitarian Law* (CUP, 2005) 48–49.

ex ante assessments, i.e. the expectations and anticipations of those in charge of the attacks;[54] it also shows that proportionality analysis is made of three steps: (*a*) collateral damage estimation; (*b*) military advantage assessment; (*c*) determination of excessiveness. Conversely, norms on proportionality offer only little guidance as to the way this analysis should be carried out in concrete cases, having particular regard to how collateral damage and military advantage should be values, and mostly how excessiveness is to be measured.

The identification of what constitutes collateral damage is dependent on the scope of the principle of distinction: harm to civilians and civilian objects, whose intentional causation is prohibited under the principle of distinction, qualifies as collateral damage when it ensues from an attack directed against a legitimate target.[55] Proportionality analysis, however, additionally requires the (metaphorical) attribution of a 'weight' to the expected collateral damage in order to compare it with the anticipated military advantage. It should be noted, in this regard, that, while it is true that provisions on proportionality do not explicitly graduate collateral damages, some distinctions can be drawn as a matter of course. It is all too clear, for instance, that damages to property weigh less than human casualties; that recoverable injuries should be valued less than permanent ones; that losses of life are more serious collateral damage than injuries. Also, it can be maintained that harm to certain categories of persons and objects is to be accorded greater weight in proportionality calculations. This could be the case of vulnerable persons (e.g. children),[56] as well as of objects enjoying special protection from attack (e.g. cultural property).[57]

The definition of the other term of the proportionality equation also has a clear connection with the principle of distinction: the notion of 'military advantage' is indeed crucial to establish whether an object

54 Yoram Dinstein, *The Conduct of Hostilities Under the Law of International Armed Conflict* (4th edn, CUP, 2022) 181–182.

55 Kenneth Watkin, 'Assessing Proportionality: Moral Complexity and Legal Rules' [2005] 8 *Yearbook of International Humanitarian Law* 3, 8–9.

56 Dinstein (n 53) 184. Such a differential treatment seems to be warranted by the fact that vulnerable persons benefit from enhanced protection under international law.

57 Henderson (n 49) 192–193. A category of object that is expressly recognized greater weight in proportionality analysis is that envisaged by Art 56 AP I ('Protection of works and installations containing dangerous forces').

may be legitimately targeted. Yet, the application of this notion differs in the two contexts, as 'military advantage" constitutes 'a threshold determination for military objective' in relation to distinction, 'and a balancing for proportionality'.[58] Unlike distinction, in other words, the principle of proportionality does not content itself with the ascertainment of the existence of a military advantage, but requires that this advantage be assigned a value, to be weighed against the expected incidental harm to civilians and civilian objects. This complicates the (already complex) assessment of the military advantage, in that it brings into play factors that are not relevant (or are relatively relevant) when the principle of distinction is at stake. In this respect, mention should be made of the actual likelihood of success of the planned attack. Indeed, the uncertainty of obtaining the anticipated military advantage 'diminishes' its value for the purposes of proportionality (while it remains immaterial in relation to issues of distinction).[59] Furthermore, the value of a military objective may vary as the conflict goes on and even during the same attack, especially if it is of large scale. This means that 'a target folder cannot assign a certain value to the military advantage of a target once and for all',[60] as the assessment of the value of attacking a certain target must be constantly updated.

Once expected collateral damage and anticipated military advantage are assessed and attributed a value, the attacking force has to establish whether the former is 'excessive' in relation to the latter. Here lies the core of the principle of proportionality, as well as the source of the most intractable problems that it poses. The problem is that what proportionality ultimately requires is a comparison 'between unlike quantities and values'[61] (e.g. between the destruction of a bridge used by enemy tanks and a certain number of civilian casualties) –an 'apples vs. oranges' dilemma, as it is often said,[62] which makes it particularly

58 Joseph Holland, 'Military Objective and Collateral Damage: Their Relationship and Dynamics' [2004] 7 *Yearbook of International Humanitarian Law* 35, 54.

59 Laurent Gisel, *The Principle of Proportionality in the Rules Governing the Conduct of Hostilities under International Humanitarian Law*, International Expert Meeting 22–23 June 2016 – Quebec (ICRC, 2018) 18.

60 Henderson (n 49) 202.

61 William J Fenrick, 'The Rule of Proportionality and Protocol I in Conventional Warfare' [1982] 98 *Military Law Review* 91, 94. See also OTP Report (n 21) para 48.

62 See, among many others, Henderson (n 49) 228–229.

challenging to identify in advance clear-cut, general rules to help the decision-maker to act lawfully. It is noteworthy, in this respect, that the examples usually provided to illustrate breaches of proportionality are pretty obvious (e.g. the bombing of an isolated fuel tanker in the middle of a densely populated city),[63] while the 'hard cases', a label which covers the vast majority of proportionality scenarios, are considered to be resolved on the basis of a contextual, case-by-case analysis.[64]

This is not to mean, however, that the combatants' discretion in this regard is left 'unfettered'.[65] Following an early intuition by William J. Fenrick,[66] courts, academics, NGOs, and military experts have been progressively converging upon the standard against which the proportionality of an attack must be assessed, that of the 'reasonable military commander'.[67] This standard is aimed at fixing the outer limits of the discretion enjoyed by military decision-makers[68] or, to put it differently, to create a zone of reasonable discretion within which all decisions about proportionality are lawful.[69] Even if objective in character, the standard of the 'reasonable military commander' remains quite elusive. The most one can say about it is that 'there will be many cases where reasonable military commanders will agree that the injury to noncombatants or the damage to civilian objects was clearly disproportionate to the military advantage gained'.[70]

3.1. Can AWS and DSS Autonomously Make Proportionality Assessments?

The principle of proportionality introduces a further layer of tasks that AWS and DSS should be able to perform, which come on top of those exacted by the principle of distinction. In the first place, these systems

63 See Gisel (n 58) 53–54.

64 *Public Committee Against Torture in Israel v Government of Israel*, HCJ 769/02 [2006] (Israeli Supreme Court) para 46.

65 Dinstein (n 53) 183.

66 Fenrick (n 50) 127.

67 See, also for references, Ian Henderson and Kate Reece, 'Proportionality under International Humanitarian Law: The "Reasonable Military Commander" Standard and Reverberating Effects' [2018] 51(3) *Vanderbilt Journal of Transnational Law* 835, 845.

68 Gisel (n 58) 64.

69 *Beit Sourik Village Council v. The Government of Israel*, HCJ 2056/04 [2004] (Israeli Supreme Court) paras 42, 46.

70 OTP Report (n 21) para 50.

should be able to make reliable predictions as to the collateral damage casualties expected from a given course of action. In fact, this does not come across as an insurmountable challenge: for more than a decade, the US Army has been using software programs to make collateral damage estimations with a view to identifying the least harmful targeting option.[71]

Far more daunting is the challenge posed by the appraisal of excessiveness. To ensure compliance with this principle, indeed, law-abiding AWS and DSS should be able to assign a value to the expected collateral damage and the military advantage anticipated and to choose a targeting option that strikes a balance between them, in a way that a reasonable military commander would find acceptable. Still, if there is one point on which most (if not all) IHL experts agree it is that, despite what the terms 'proportionality', 'calculus', and 'equation' would suggest, this balancing exercise does not lend itself to be straightforwardly captured by a 'mathematical formula'.[72] It is no surprise, therefore, that these authorities are often cited in support of the conclusion that the proportionality principle is unsuitable to be coded into a computer algorithm.[73] Indeed, the very terms generally employed in legal scholarship to describe the kind of judgements involved in proportionality analysis – 'reasonableness', 'common sense', 'good faith'– are simply a 'non-starter' for programmers.[74]

The reasons for this 'translatability' problem have been well explained by Andreas Matthias.[75] Matthias observed that the idea of a software programmed to inhibit unethical courses of action by weapons systems moves from the assumption that it is possible (*a*) to assign to every

71 Thurnher (n 45) 221; Sharkey (n 27) 788–789.

72 Michael Wells-Greco, 'Operation 'Cast Lead': *Jus in Bello* Proportionality' [2010] 57(3) *Netherlands International Law Review* 397, 416.

73 See, for instance, Jeroen van den Boogaard, 'Proportionality and Autonomous Weapons Systems' [2015] 6(2) *Journal of International Humanitarian Legal Studies* 247, 283; Tanya Krupyi, 'Of Souls, Spirits and Ghosts: Transposing the Application of the Rules of Targeting to Lethal Autonomous Robots' [2015] 16(1) *Melbourne Journal of International Law* 1, 28.

74 Maziar Homayounnejad, '*Ensuring Lethal Autonomous Weapon Systems Comply with International Humanitarian Law*' (TLI Think! Paper 85/2017) 25. See also Suresh Venkatasubramanian, 'Structural Disconnects between Algorithmic Decision-Making and the Law' (*Humanitarian Law & Policy*, 25 April 2019) < https://blogs.icrc.org/law-and-policy/2019/04/25/structural-disconnects-algorithmic-decision-making-law> accessed 23 April 2024.

75 *Ibid* 9.

possible action 'a scalar 'morality' value'; (*b*) to measure this value 'with precision at any point in time'; (*c*) to compare this 'morality value' to that 'of other, alternative actions'; (*d*) to identify, for every moral problem, 'a single optimal target value', viz. the 'one and only one optimal solution', with the consequence that 'all alternative paths of action [would be] deviations from [it]'.[76] Still, moral 'problems', including those involving the use of lethal force on the battlefield, are so labelled for a reason, namely that none of them 'comes with a 'reference answer' which can be considered to solve the moral dilemma in a generally satisfactory way'.[77] If this is true, then the idea of programming an algorithmic inhibitor of immoral actions loses its very *raison d'être*, for '[w]here there is no target value, we cannot speak of a regulation process'.[78]

While originating from the realm of philosophy, Matthias' objection holds true in relation to the IHL principle of proportionality. Proportionality, indeed, does not provide 'reference answers': multiple courses of action are available and none of them may be said to be completely satisfactory, in that each involves a certain amount of harm to civilians and/or civilian objects, whose acceptance is an ultimately moral decision.[79]

Some scholars attempted to dodge this problem by arguing that it is always possible to ensure that the proportionality of the attack is in fact appraised by human beings, by 'pre-determin[ing] the upper limit for civilian casualties and various types of property damage prior to authorising the AWS mission'.[80] A similar argument could be made in relation to DSS to affirm that it is enough, legally speaking, they are designed to correctly identify the targets and estimate collateral damage, since proportionality calculations will still be made by human commanders or operators.

---

76 International Committee of the Red Cross, *Ethics and Autonomous Weapon Systems: An Ethical Basis for Human Control?* (CCW Group of Governmental Experts on Lethal Autonomous Weapons Systems, 28 March 2018, UN Doc. CCW/GGE.1/2018/WP.5) para 21.

77 *Ibid.*

78 *Ibid.*

79 David Kennedy, *Dark Side of Virtue: Reassessing International Humanitarianism* (Princeton University Press, 2004), 290–291; Dale Stephens and MW Lewis, 'The Law of Armed Conflict – A Contemporary Critique' [2005] 6 *Melbourne Journal of International Law* 55.

80 Ian Henderson, Patrick Keane and Josh Liddy, 'Remote and Autonomous Warfare Systems: Precautions in Attack and Individual Accountability', in JD Ohlin (ed.), *Research Handbook on Remote Warfare* (Edward Elgar Publishing, 2017) 335, 352.

At first sight, this line of reasoning may come across as reassuring, in that it seems that the gist of proportionality assessments is kept firmly in human hands. Yet, one should beware that the pre-determination of acceptable collateral damage should not be conducive to a trivialization of proportionality assessments, by turning it into a merely quantitative analysis.[81] This means that human operators cannot limit themselves to putting a 'price tag', in terms of civilian harm, on each generic set of targets (e.g., by 'valuing' the destruction of an enemy tank according to the number of permissible civilian deaths).[82] The authorization of an AWS sortie, or of an attack recommended by a DSS, should instead be made conditional upon a careful ponderation of the expected collateral damage, which takes into account the relative weight of the categories of civilians and civilian objects that may be incidentally affected. Only on the basis of such a comprehensive assessment would the human commander/operator be in a position to appraise whether the military advantage sought by the AWS mission (or the DSS' recommendation) is sufficiently important to justify the expected collateral damage.

The humanitarian catastrophe of the ongoing conflict in Gaza provides a dramatic illustration of the problematic implications of using AI for target identification, couples with a blanket 'price tag' approach to proportionality calculations. According to some sources, in the early weeks of the conflict the number of civilian casualties permitted for the killing of each junior Hamas operative, including those identified by the abovementioned Lavender system, was set at 20. It is no coincidence, on the other hand, that a concept closely related to 'price tagging', the 'non-combatant casualty cut-off value' (NCV), is no longer included in the official US military doctrine.[83]

81 Markus Wagner, 'The Dehumanization of International Humanitarian Law: Legal, Ethical, and Political Implications of Autonomous Weapon Systems' [2014] 47 *Vanderbilt Journal of Transnational Law* 1371, 1397–1398.

82 Louise Doswald-Beck, 'The Value of the 1977 Geneva Protocols for the Protection of Civilians', in MA Meyer (ed.), *Armed Conflict and the New Law: Aspects of the 1977 Geneva Protocols and the 1981 Weapons Convention* (BIICL, 1989) 153, 156 ('it is impossible to state that a factory is worth x civilians'). Regrettably, this view has seemingly been put forward by influential commentators like Schmitt and Thurnher (n 43) 256. For a more nuanced approach, although still based on a 'price-tagging' approach, see Maciek Zając, 'AWS compliance with the ethical principle of proportionality: three possible solutions' [2023] 25(1) *Ethics and Information Technology* 1, 13–14.

83 US Department of Defense, *Civilian Harm Mitigation and Response Action Plan* (25 August 2022).

But what about the possibility to let AWS and DSS make proportionality calculations on their own? If it is true that the principle at hand could not be 'translated' in 'a set of decision rules', researchers are exploring the possibility that AWS and DSS develop a 'practice', built on experience and training, on whose basis to perform legally acceptable proportionality calculations.[84] To put it in the simplest terms, since the principle cannot be directly coded, military research could focus on conceiving of a learning artificial architecture allowing AWS and DSS to approximate legally acceptable proportionality decisions, on the basis of both pre-deployment training and battlefield experience. At the training stage, in particular, the system would be fed with a dataset containing (historical or fictional) warfare scenarios where proportionality issues were resolved in line with the standard of the 'reasonable military commander', *at least in the opinion of the instructors*. This would allow the system, confronted with a new proportionality scenario, to 'retrieve' from its database the solution of the case that is most similar or to develop its own way to make reasonable proportionality decisions.[85]Instructors will continue to correct the systems' mistaken decisions and withhold deployment on the battlefield, until they are satisfied with the level of IHL-compliance of the system's proportionality assessments.[86]

It is worth reiterating that we are not talking about imminent technological innovations, but about long-term goals, whose achievement may be plausibly located in a distant future. However, as some military research projects have taken up this challenge, it is worth pushing our investigation a bit further and testing the claim, often made by AWS proponents, whereby autonomous performance of proportionality assessments would be beneficial to the respect of IHL. In particular, reference is made to the expectation that AWS (and DSS) will one day outperform humans in complying with IHL principles, including that of proportionality. This may be possible with respect to distinction and certain aspects of proportionality (such as the estimation of collateral damage), but it is doubtful that it will ever happen in relation to the core of proportionality,

84 Ronald C Arkin, *Governing Lethal Behavior in Autonomous Robots* (CRC Press, 2009) 47.

85 The latter would be the case if a reinforcement learning technique is opted for (International Committee of the Red Cross, *Autonomy, Artificial Intelligence and Robotics: Technical Aspects of Human Control* (ICRC, August 2019) 16); the former if a Case-Based Reasoning method is adopted (Arkin (n 83) 47–48).

86 Backstrom and Henderson (n 34) 494.

i.e. balancing. To the extent that the principle of proportionality does not set out an objectively identifiable goal to be achieved (such as 'zero civilian casualties'), AWS' and DSS' training will be heavily influenced by the instructors' personal beliefs about what a reasonable military commander would do. The best that can be wished for, thus, is for these systems to carry out proportionality analysis as well as their instructors, who, it can be hoped, but not guaranteed, will be conscientious and competent experts.

3.2. Summary of the Discussion

Unlike distinction, the principle of proportionality cannot be captured by a formula prescribing, in general and abstract terms, the course of action required by the norm. This unique feature of the principle makes it particularly unsuitable to be 'translated' into a computer program.

To overcome this problem, some authors suggest to pre-determine proportionality calculations at the planning stage, so as to ensure that they are made by human commanders. This may provide an adequate legal solution in certain scenarios, but one should carefully avoid turning predetermination into an oversimplification of proportionality assessment.

Autonomous proportionality assessment is in principle possible, although not in the immediate future, thanks to the developments in machine learning techniques. However, expectations about the humanitarian benefits of this should be significantly lowered.

4. The Third Prong of the Test: Precaution

The principles of distinction and proportionality may be described as envisaging obligations that are negative in character: do not target protected persons and objects, do not launch disproportionate attacks. IHL, however, goes beyond that by requiring States to comply with certain positive obligations, which are instrumental to the fullest adherence to the prohibitions set forth by the principles of distinction and proportionality. Here lies the core of the principle of precaution, which brings added normative value to the principles of distinction and proportionality, by introducing the duty take affirmative action, to 'do everything feasible' (or 'take all feasible precautions') to ensure that these principles are fully complied with.[87]

87 Art 57 AP I.

Precautions cover a wide range of measures of very different character (e.g. the issuance of advance warning). For the purposes of our discussion, it will suffice to focus on two of them. On the one hand, there is the obligation to undertake all practicable efforts 'to obtain reliable information' as to whether the target is a legitimate military objective.[88] On the other hand, everything feasible should be done to ensure that combatants have an adequate control over the effects of the weapons they use, with a view to avoiding or at least minimizing damages to civilians and civilian objects.[89] States must therefore mobilize all assets at their disposal, where reasonably available, to facilitate correct target identification. This requirement has to be understood as including human resources too, wherever they are conducive to more accuracy.[90] Also, if practicable, combatants must be in a position to control and limit the effects of the weapons they employ.

4.1. The Normative Implications of the Principle of Precaution for Human-Machine Interaction on the Battlefield

Relying on the research goal of designing AI-based systems outperforming humans on the battlefield, some scholars have argued that autonomy in weapons systems would make 'additional precautions possible', with the consequence that a commander having this kind of weapon in his/her arsenal would be legally required to use it instead of a human-controlled one, wherever feasible.[91]

This argument cannot be accepted on its own terms. This is not so much because, as we have tried to show above, expectations about the performance of AI-based systems on the battlefield should be significantly lowered. Rather, it is because the removal of the human element from the targeting process would preclude a feasible precautionary measure capable of preventing violations of the principles of distinction and proportionality.

88 Program on Humanitarian Policy and Conflict Research at Harvard University, *Manual on International Law Applicable to Air and Missile Warfare* (CUP, 2013) 144 (HPCR Manual).

89 Henckaerts and Doswald-Beck (n 52) rules 17-19.

90 Marina Mancini, 'Air Operations against the Federal Republic of Yugoslavia (1999)', in Natalino Ronzitti and Gabriella Venturini (eds.), *The Law of Air Warfare: Contemporary Issues* (Eleven 2006) 273, 278.

91 See, for instance, Peter Margulies, 'The Other Side of Autonomous Weapons: Using Artificial Intelligence to Enhance IHL Compliance', in RTP Alcala and Eric Talbot Jensen (eds.), *The Impact of Emerging Technologies on the Law of Armed Conflict* (OUP, 2019) 147.

In this regard, one should preliminarily note that, even the most enthusiastic proponents of AI-based systems are ready to admit that they will never be error-free.[92] Yet, this admission is usually accompanied by the observation that human combatants are far from perfect either;[93] that international law would not exact from AI more than it does from human war fighters; and that in any case autonomous targeting will ensure higher standards of compliance.[94] Still, regardless of whether one accepts these arguments, they are ultimately immaterial to our discussion. Indeed, they all proceed from the idea that the legal performances of human and autonomous targeting should be compared along a linear scale to determine which one is the best in terms of IHL-compliance. However, such an approach would only be warranted if human combatants and machines were in principle prone to the same flaws, biases and vulnerabilities, since this would entail that, once humans are outperformed by autonomous weapons, they cannot add any value to the legality of the targeting process.

But this is not the case. A growing body of studies in AI shows that, even when machines prove better than humans in a given task, they still incur mistakes that a human would never have made and for reasons that are sometimes obscure to their programmers themselves.[95] Moreover, machines can be fooled into badly wrong assessments with tricks (e.g. a minor adjustment of an image) that would not even be noticed by humans (so-called 'adversarial attacks')[96] and, conversely, they can 'cheat' human evaluators, by learning how to get their approval during the training phase, regardless of whether they are pursuing the assigned task correctly.[97] In other words, human and autonomous

92 See, for instance, Arkin (n 83) 30–31.

93 *Ibid* 31 – 36; Marco Sassóli, 'Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to be Clarified' [2014] 90 *International Law Studies* 308, 310.

94 Arkin (n 83) 29-30.

95 For an updated survey of – sometimes genuinely weird – AI failures, see Aaron Drapkin, 'AI Gone Wrong: An Updated List of AI Errors, Mistakes and Failures' (*tech.co*, 18 April 2024) <https://tech.co/news/list-ai-failures-mistakes-errors> accessed 23 April 2024.

96 Makaela Nartker, Zhenglong Zhou, and Chaz Firestone, 'When will AI misclassify? Intuiting failures on natural images' [2023] 23(4):4 J*ournal of Vision* 1.

97 Mark Farragher, 'This clever AI Hid Data from its Creator to Cheat at its Appointed Task' (*Medium*, 14 March 2019) <https://mdfarragher.medium.com/this-clever-ai-hid-data-from-its-creator-to-cheat-at-its-appointed-task-8b91ed9872f5> accessed 23 April 2024.

decision-making processes remain qualitatively different and are likely to err in qualitatively different ways.

As explained in a 2019 ICRC report, the source of such qualitative difference lies in what AI experts call the 'semantic gap'. This term indicates the fact that machines do not 'see' the world in the same way as humans, as only the latter have a proper understanding of the meaning or concept of the objects they classify and recognize, and of the context in which they are placed.[98] Accordingly, while it is reasonable to expect that the overall performances of AI-based systems will one day surpass those of humans (at least in some respects), this will not necessarily be the case *in each and every situation* where they are deployed. Indeed, even the most sophisticated system could commit (or be induced to commit) mistakes that could have been avoided had a human operator been involved in the decision-making loop. If we consider that, in the context of targeting, 'mistakes' generally amount to unlawful harm to civilians or civilian objects – that is, precisely what precautionary measures are meant to prevent – it becomes clear that it makes little sense to maintain in the abstract that autonomous targeting would be normatively preferable under the principle of precaution.

As we have said, the principle of precaution requires States to do everything feasible to facilitate correct target identification and to avoid (or minimize) collateral damage. The wide breadth of the obligations imposed by this principle (expressed by the use of words such as 'everything' or 'all') does not substantiate an all-or-nothing approach, whereby the critical functions of target selection and engagement should be entrusted either entirely to machines or entirely to humans. Rather, to ensure respect for the dictate that '*everything feasible*' is done, States must focus on developing human-machine relationships that bring out the best in both, or –which is the same thing– compensate for their mutual weaknesses.[99] Only in this way will they be able to exploit the full potential of the assets at their disposal in order to prevent or minimize unlawful harm.

98 ICRC (n 84) 20.

99 Noel E Sharkey, 'Staying the Loop: Human Supervisory Control of Weapons', in Nehal Bhuta et al., *Autonomous Weapons Systems: Law, Ethics, Policy* (CUP, 2016) 23. This idea is also embodied in the concept of 'centaur warfighting' advocated by Paul Scharre (Paul Scharre, 'Centaur Warfighting: The False Choice of Humans vs. Automation' [2016] 30(1) *Temple International and Comparative Law Journal* 151.

Awareness of the above has led to a number of doctrinal attempts to craft an ideal human-machine interaction, sometimes explicitly characterizing it as a precautionary measure.[100] These debates fall outside of the scope of the present Chapter, as they relate to the overarching concept of Meaningful Human Control, which is the object of several in-depth treatises.[101] For the purposes of the present discussion, we may observe that, in order to be properly qualified as a precautionary measure, this interaction must be designed in such a way as to comply with two guidelines.

On the one hand, the involvement of a human operator must add value to the targeting process.[102] In the debate on AWS, the risk has often been emphasized that the human contribution could be nullified by the so-called 'automation bias', i.e. the propensity to prefer solutions elaborated by the machine, even when contradicted by information coming from other sources.[103]

A tragic example in this sense is represented by Patriot missiles' fratricide incidents, which occurred in Iraq in 2003. The MIM-104 Patriot is an air defense system designed to intercept and destroy tactical ballistic missiles, cruise missiles and advanced aircrafts, which is manufactured by the US defense contractor Rayhteon.[104] The system autonomously identifies possible ballistic missile targets based on an assessment of the speed, altitude, and behavior of the objects detected by its radar, but leaves the actual decision to engage the targets to human operators. It was deployed during *Operation Iraqi Freedom* in 2003, resulting in a

100 This is the case, for instance, of Robin Geiss and Henning Lahmann, 'Autonomous Weapons Systems: A Paradigm Shift for the Law of Armed Conflict?', in Ohlin (n 76) 371, 397. See also International Panel on the Regulation of Autonomous Weapons (iPRAW), *Focus on Human Control*, Report No 5 (Berlin, August 2019) 11 (adhering to the view whereby 'the principle of precaution inherently requires some sort of human control during the attack because the abstract target selection criteria do not allow for sufficient legal assessments prior to the attack'); Jeffrey S Thurnher, *Means and Methods of the Future: Autonomous Systems*, in PAL Ducheine, MN Schmitt and FPB Osinga (eds.), *Targeting: The Challenges of Modern Warfare* (Springer, 2016) 177, 190 (arguing that '[i]n many circumstances, a manned or remotely controlled system may prove to be the necessary choice' to comply with the principle of precaution under IHL

101 See, most recently, Giulio Mecacci et al (eds.), *Research Handbook on Meaningful Human Control of Artificial Intelligence Systems* (Edward Elgar Publishing 2024).

102 Backstrom and Henderson (n 34) 496.

103 LJ Skitka, KL Mosier and Mark Burdick, 'Does Automation Bias Decision-making?' [1999] 51(5) *International Journal of Human-Computer Studies* 991.

104 See 'Patriot Missile Long-Range Air-Defence System, USA' (Army Technology)<https://www.army-technology.com/projects/patriot> accessed 23 April 2024.

very high success rate against Iraqi missiles. On two occasions, however, Patriot misidentified friendly aircrafts as enemy missiles, with the consequence that a British Tornado and a Navy F-18 fighter jet were shot down and their crews killed.[105] Crucially enough, for our purposes, subsequent Army investigations found that the ultimate cause of the fratricide incidents was a widespread culture, among Patriot operators, of 'unwarranted and uncritical trust in automation'.[106] The same 'uncritical trust' seems to characterize the use of Lavender in the Gaza conflict. In this regard, it has been reported that 'if Lavender decided an individual was a militant in Hamas, they were essentially asked to treat that as an order, with no requirement to independently check why the machine made that choice or to examine the raw intelligence data on which it is based'.[107]

Leaving aside the blatant absurdity of a practice that *de facto* equates a recommendation by a DSS with a superior's order, one should note that, contrary to what is sometimes argued,[108] the human propensity to trust machines is not in itself a valid reason to exclude the legal relevance of keeping human operators in the decision-making loop altogether. Rather, it urges the adoption of additional measures to guarantee that the human-machine interaction is effective.

In the first place, human operators should be provided with full knowledge of the system's capabilities and limits, with a view to educating them to trust the technology without completely relinquishing giving up their judgement and critical sense.[109] Moreover, the human-machine distribution of control privileges should be conceived of in such a way as to counter the risk of automation bias. It should be noted, in this respect, that a relationship in which the human chooses a course of action from a list of alternatives provided by the system is less conducive to such a

105 JK Hawley, *Patriot Wars: Automation and the Patriot Air and Missile Defense System* (Center for a New American Security, January 2017) 7-8.

106 Excerpts from the report are in JK Hawley, 'Not by Widgets Alone: The Human Challenge of Technology-intensive Military Systems' (*Armed Forces Journal*, 1 February 2011) <http://armed-forcesjournal.com/not-by-widgets-alone/> accessed 23 April 2024.

107 Abraham (n 2).

108 Sassóli (n 92) 337.

109 This point is well underscored by Paul Scharre who, on the basis of an insightful comparison of the approaches followed by the users of the Patriot and Aegis systems, indicated the latter as a virtuous model in that the machine effectuates warfighters' decision-making, instead of merely replacing it. Scharre (n 26) 161–173.

bias than one in which the weapon recommends a targeting decision, and the human authorizes the attack.[110] And both undoubtedly preserve human awareness better than a distribution of control privileges where autonomous targeting is subject only to human override.[111]

On the other hand, a proper human-machine relationship should address the weaknesses of the human element as well. Especially when put under stressful conditions, humans can be induced to misinterpret the data available to them 'in an unconscious attempt to make available evidence fit a preconceived scenario'.[112] This bias –commonly known as 'scenario fulfillment'– was studied extensively in relation to the 1988 USS Vincennes incident, which involved the shooting down of an Iranian airliner by a US warship in the Persian Gulf during the Iran-Iraq war. The most salient feature of this incident is that, although the available data clearly indicated that the aircraft posed no threat, the crew nevertheless mistakenly believed it to be an Iranian F-14 fighter and attacked it, causing the death of all 290 people on board.

This episode is sometimes relied on to deny that human control is an effective precautionary measure.[113] With all due respect, this pushes the argument too far. While it is true that scenario fulfillment is a typically human bias, this does not automatically rule out that, in other contexts, human control may actually help to avoid harmful events. If there is a lesson to be learned from the USS Vincennes incident, it is that human judgement should not be blindly trusted as well, and that the human-machine relationship should take this into account. One might envisage a 'two-key' mechanism whereby coded authorization by two separate individuals would be required to override the system's assessment that a targeting decision would be unlawful.[114] In this way, human decision-makers would be forced to pause and reconsider their judgement in light of the system's explanations –a measure that would probably have

110 Sharkey (n 98) 34-35; Backstrom and Henderson (n 34) 496.

111 In this perspective, the US Law of War Manual's suggestion that additional precautions should be taken in relation to AWS, which include 'monitoring the operation of the weapon' (US Department of Defense, *Law of War Manual* (updated 31 July 2023), 363) is certainly a welcome step, but does not come across as the best available solution.

112 Molly Moore, 'Vincennes Report Admits Errors but Urges No Disciplinary Action' (*The Washington Post*, 20 August 1988).

113 Schmitt and Thurnher (n 43) 248.

114 Arkin (n 83) 152–153.

prevented many tragic incidents, including that of the USS Vincennes.

The foregoing may appear quite bold statements as to what States would be obliged to do under the principle of precaution. In fact, it should always be borne in mind that the scope of the obligations deriving from the principle of precaution is subject to the limit of feasibility. As a consequence, while it is normatively desirable that States adopt all the above measures, they are excused for not taking them if they are wholly or partly impracticable. Some of the above requirements come across as relatively unproblematic. This is certainly the case of the obligation to provide adequate training that, as a matter of principle, should always be considered feasible. Likewise, measures aimed at ensuring the double-checking of human decisions could be more or less challenging for AI research but, once the necessary technology is developed, their implementation should not be hindered by military considerations.

Things stand quite differently as to the human-machine distribution of control privileges: there could be indeed scenarios where the enactment of higher levels of human control would be unfeasible because of competing military considerations, insofar as it would pose serious security risks (for defensive systems) or it would hinder mission accomplishment (for offensive systems). A relatively straightforward example of the former case may be that of an air defense system, such as the US Aegis Ballistic Missile Defense System or the Israeli Iron Dome, which has to cope with a saturation attack where multiple threats are incoming at a very fast pace. In this hypothesis, ensuring that human operators have full control over every single targeting decision would be prohibitive, and even the exercise of veto powers would be to a large extent an unavailable option. Accordingly, the decision to lower the level of human control (or even to exclude it, in the most extreme circumstances) would not be incompatible with the requirement of precaution.

More controversial is the case where human involvement in the decision loop –although technically practicable– could hinder mission accomplishment. Unlike the defensive scenario above, in this case the commander has a choice: (s)he can decide not to carry out the mission. In hypotheses of this kind, a crucial role will be played by humanitarian considerations, which also have to be factored into feasibility analysis.

If, given the features of the operational environment, the risks for protected persons and objects are virtually absent, the required level of human control may be lowered. Should such risks be instead present – because, for instance, the mission has to be accomplished in a cluttered urban warfare scenario– the principle of precaution will require the commander to maintain a higher level of human control and, if this makes mission accomplishment too hard, to cancel the attack and opt for a different course of action.[115]

4.2. Summary of the Discussion

Contrary to what is sometimes argued, exclusive reliance on AI systems in warfare –however sophisticated these technologies may be– is not in itself a mandatory precautionary measure under IHL. The obligation to do 'everything feasible' to prevent unlawful harm, indeed, requires States to develop human-machine relationships that bring out the best in both machines and humans, by compensating their reciprocal weaknesses, and not to give absolute preference to one over the other.

To be an effective precautionary measure, the human-machine interaction should be shaped so as to counter the risk of automation bias and scenario fulfillment. The former should be addressed by providing adequate training to human operators, and by choosing a level of human control that preserve human awareness and critical sense; the latter, by allowing the weapon to suspend human targeting decisions that it deems unlawful and request a double-check.

The obligation to adopt the above measures is conditional upon their feasibility. This is particularly true of human control, because in certain scenarios it could hinder mission accomplishment or pose serious security risks, making it impracticable –at least from a purely military perspective. Whether this justifies reducing or even excluding human control cannot be assessed solely on the basis of military evaluations, though. Humanitarian considerations (e.g. the risk of civilian casualties) also play a role in the feasibility equation, balancing the need to achieve the mission's goal.

115 WH Boothby, 'Some Legal Challenges Posed by Remote Attack' [2012] 94 (886) *International Review of the Red Cross* 579, 584.

5. Conclusions

It is now time to wrap up the long discussion carried out so far and to formulate a tentative answer to the questions raised in the introductory section:

- Are AWS and DSS, in their normal or intended circumstances of use, capable of complying with the applicable distinction obligations?
- Are AWS and DSS, in their normal or intended circumstances of use, capable of making the proportionality assessments required by IHL?
- Do AWS and DSS, in their normal or intended circumstances of use, hinder compliance with the obligation to take every feasible precaution to avoid (or minimize) civilian harm?

The first prong of the test presents an eminently technological challenge. Distinction obligations under IHL require that AWS and DSS are able to properly recognize and assess the elements and the dynamics that, in a given operational environment, are relevant for the lawfulness of a targeting decision (what we have referred to as SA capabilities). The scale of the technological challenge posed by this requirement is not the same for every possible use of AWS and DSS, but is determined –individually or in combination– by the following variables: (*a*) the features of the operational environment (cluttered vs. uncluttered); (*b*) the targeting mode (dynamic vs. deliberate); (*c*) the nature of the target (combatants and inhabited military objects vs. uninhabited military objects).

Current technology is up to the task only in a limited number of circumstances, where the values of the above variables converge towards lowering the required level of situation awareness. These may include air defense systems, which are entrusted with a relatively simple distinction task (detecting incoming threats), generally operate in uncluttered environments, and do not have to recognize messages of surrender from potential targets; fire-and-forget munitions that are launched in uncluttered scenarios; loitering munitions that are activated in uncluttered scenarios and are programmed to attack uninhabited military objects (e.g. enemy drones). Similar considerations apply to DSS' capability of proposing lawful targeting decisions. *In relation to these hypotheses, thus, the first prong of the test may be considered to have been passed even at the current state of technological development.*

In the vast majority of cases, however, the lawful use of AWS, as well as the ability of DSS to recommend legitimate targeting decisions, will depend on whether the system will be capable of autonomously acquiring higher levels of situation awareness. There is still a long way to go in AI research in this regard, and it is not possible to make reliable predictions about whether and when this goal will be achieved. *In relation to these hypotheses, the first prong of the test can therefore be considered to have failed, but this failure depends on the state of technological development.*

The second prong of the test raises instead a translational challenge. The principle of proportionality –and notably of the balancing exercise lying at its cor – seems ill-suited to being translated ('coded') into a computer program. As we have seen, two options have been put forth to address this problem. On the one hand, proportionality calculations could be anticipated at the planning stage, so as to ensure that they are still carried out by human decision-makers. This may be an appropriate legal solution in the case of deliberate targeting, but care should be taken not to turn pre-determination into an oversimplification of proportionality assessment. *Subject to these conditions, and provided that the required level of situation awareness is met, the second prong of the test may hence be considered to have been passed when proportionality calculations are made by a human being at the planning stage.*

On the other hand, AWS and DSS might one day be programmed –thanks to reinforcement machine learning techniques– to perform proportionality calculations on their own, not on the basis of some sort of algorithmic formula (which would be impossible to determine in advance), but by directly emulating human reasoning. Yet, such a technological development, although sometimes discussed in the literature, does not appear to be achievable in the near future. *Accordingly, if proportionality calculations are not made by humans, the second prong of the test can be considered to have failed, at least on the basis of the current state of technological development.*

The third and final prong of the test contains a *design* challenge. The positive obligation to take precautionary measures implies that States must develop human-machine relationships that bring out the best in both machines and humans by compensating for their reciprocal weaknesses. On this assumption, a number of specific features of such

a human-machine relationship have been distilled, the most important of which is –for the present discussion– *the need to preserve the operators' awareness and critical sense by ensuring appropriate levels of human control.* The duty to adopt this measure is subject to two qualifications. On the one hand, the obligation to take precautionary measures is instrumental to the fulfillment of the requirements deriving from the principles of distinction and proportionality. Accordingly, it arises only when and to the extent that issues of distinction and proportionality are at stake. On the other hand, precautionary measures need to be adopted only if they are militarily feasible. This implies that, in principle, human control could legitimately be limited or excluded when it could pose serious security risks (e.g. because it hinders a timely reaction against a saturation attack).

Notwithstanding these qualifications, this prong of the test undoubtedly remains the most demanding. Unlike the other two prongs, it does raise a challenge to full autonomy that holds true even with regard to futuristic systems capable of carrying out reliable distinction and proportionality analysis in any warfare scenario. This is an important point, as it partially disproves the conventional wisdom that IHL-based objections to complete reliance on AI technologies in warfare are technologically contingent.[116]

In this perspective, the removal of the human element from the critical functions of target selection and engagement would be legally acceptable only if (*a*) the system is used in a context where issues of protection of civilians under the principles of distinction and proportionality do not arise; or (*b*) human control would be impracticable because it would pose serious security risks. *Subject to these conditions, the third and final prong of the test may be considered satisfied.*

The foregoing calls for a serious reconsideration of the appropriateness of a legal regulation of these systems aimed at identifying a one-size-fits-all solution. The vast taxonomy of AWS and DSS and their diverse applications dictates the adoption of a more nuanced approach. More specifically, a number of variables have been detected that warrant

116 International Committee of the Red Cross, *Ethics and Autonomous Weapon Systems: An Ethical Basis for Human Control?* (CCW Group of Governmental Experts on Lethal Autonomous Weapons Systems, 28 March 2018, UN Doc. CCW/GGE.1/2018/WP.5) para 21.

differential legal treatment depending on the characteristics and intended uses of the system. This highlights the need for a more tailored and context-sensitive approach to regulating the use and deployment of these technologies, particularly with regard to the crucial issue of human-machine interaction.[117] But this, of course, goes well beyond the scope of the present inquiry.

117 The importance of this issue has been acknowledged by the Group of Governmental Experts on Lethal Autonomous Weapons Systems. See Guiding Principles affirmed by the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System, UN Doc. CCW/MSP/2019/9, Principle (c).