Designing International Humanitarian Law into military autonomous devices

Jonathan Kwik[1][0000−0003−0367−5655], Tomasz Zurek[2][0000−0002−9129−3157]*, and Tom van Engers[3][0000−0003−3699−8303]

[1] Faculty of Law, University of Amsterdam
[2] T.M.C. Asser Institute, R.J. Schimmelpennincklaan 20-22 2517 JN The Hague
[3] Complex Cyber Infrastructure, Informatics Institute, University of Amsterdam
h.c.j.kwik@uva.nl,t.zurek@asser.nl, T.M.vanEngers@uva.nl

Abstract. This position paper presents a discussion on the problem of implementing the rules of International Humanitarian Law in AI-driven military autonomous devices. We introduce a structure of a hybrid data- and knowledge-driven computational framework of a hypothetical targeting system built from the ground up with IHL compliance in mind. We provide a model and a discussion of necessary legal tests and variables.

Keywords: Military AI · International Humanitarian Law · Autonomous devices.

1 Introduction

The application of artificial intelligence (AI) in weapon systems has become a major point of contention in the past decade. While many States have embraced the potential that AI brings for increasing precision and speed, improving their warfighting capacity, and reducing unneeded casualties [46, 30], a significant opposition group has also formed which contests whether AI can ever be used in military contexts in a lawful and ethical manner [44].

One point which is not in debate is the applicability of international humanitarian law (IHL), the body of international law which governs the conduct of parties to an armed conflict. In particular, IHL [2] provides that belligerents do not have full freedom in their choice of means of warfare, i.e., the weapons they deploy. Any new weapon adopted by belligerents must be in compliance to IHL, and any new technology introduced must conform itself to these existing rules [25]. This includes AI. States [48, 6], international organisations and NGOs [24, 34] and commentators [46, 39] universally agree that any weapon system which incorporates AI must uphold IHL.

In light of this universal point of departure, there has been disagreement whether AI can actually be designed to comply with the legal rules. Many have

* Tomasz Zurek received funding from the Dutch Research Council (NWO) Platform for Responsible Innovation (NWO-MVI) as part of the DILEMA Project on Designing International Law and Ethics into Military Artificial Intelligence.

expressed doubt that this can be done, arguing that AI will never be able to replace humans in this regard [42]. Those that elaborate usually point to the many (subjective) variables involved in IHL decisions and that these legal evaluations can only practically be effectuated by humans, not by the narrow AI of today [9]. In particular, the principles of distinction and proportionality in IHL are frequently cited as examples of rules which would be impossible to implement through AI [43]. Indeed, if a system is unable to properly execute these tests but is deployed regardless for tasks which would require it to conduct such legal assessments, it would be deployed unlawfully.

In this article, we counterbalance this perspective by proposing a structure of a hypothetical system which is constructed from the ground up with IHL rules in mind. The basic structure of the system is based on a commander's targeting cycle. During this process, commanders rigorously conduct several evaluations derived from IHL, and it is one of the primary mechanisms which ensure that principles such as distinction and proportionality are upheld in the field [8]. By translating this process into an equivalent in AI form, we provide one potential way IHL can be designed directly into an AI weapon system, and demonstrate that the demands for an IHL-compliant AI weapon system can theoretically be met. Our proposal aims at filling the gap between legal research on IHL and research on AI-driven autonomous decision-making systems.

Our system is deliberately built optimistically, in the sense that we capture all relevant legal tests required during targeting directly into one system. This includes target selection and sorting, distinction, proportionality calculation, and harm minimisation. It reflects what some define as a (fully) autonomous weapon system [38], i.e., where AI takes over all the functions of a commander. However, not all AI-controlled weapon systems will *necessarily* perform all of the steps described in our framework. A decision-making aid might only need the target sorting and collateral damage calculation functions, while a smart missile might only execute the proportionality calculation functionality. We make no statements on how, in detail, particular modules would be built in practice and the feasibility of such a system [30]. Rather, our main aim is explorative, i.e., to demonstrate that purely from a programming perspective, such a task can be undertaken. Another reason for including all legal tests into one framework is that variables such as military advantage and harm reduction are utilised throughout the various steps of the cycle. By capturing the entire targeting process at once, we are able to illustrate how different legal tests interlink and draw from the same variables and inputs.

Another distinctive feature of our framework is its hybrid nature which combines knowledge-driven and data-driven reasoning. One major dilemma often raised in the debate concerning AI and weapons is the dual problem that, on the one hand, the complexity and dynamicity of the modern battlefield practically requires resort to data-driven techniques such as deep neural networks for adaptability [12, 41], while on the other hand, a level of decision-making transparency is required in IHL for the purposes of predictability and accountability

[27, 20]. Through the use of a hybrid system, we draw on the strengths of both techniques while addressing both these challenges.

While we focus exclusively on the legal duties in IHL as the basis for our framework, it should be noted that in practice, many other factors - such as political and ethical perspectives - will also be important when defining the system's design requirements. [11, 21, 16, 33, 7] provide useful overviews of such considerations in the field of AI and weapon systems. We do not integrate these factors into our system at this point to maintain the framework's generality, as each State and military organisation will have different policies in force. By focusing primarily on the legal requirements, which are universal and non-derogable, we present a framework that is at a minimum IHL-compliant, after which organisations can adopt additional ethical, organisational and political requirements in accordance with their respective preferences and policies.

This paper proceeds as follows. We begin with a brief comment on the law applicable to military systems and the targeting cycle in general. We then explore, in depth, the military targeting cycle upon which our system's framework was built. This includes two aspects. First, we discuss the formal steps of the targeting process and how these are implemented during military operations. Second, we discuss what IHL principles are relevant for the targeting stage and their respective timings. We then integrate the law into the targeting process and present formalisations of specific rules of IHL, such as proportionality and minimisation, thereby also highlighting the variables which are the most important as inputs for these tests. On this basis, we subsequently introduce the framework of our system, briefly discuss the necessary functionalities of the system, its structure, and required data.

2 The law and the operational framework

Limitations on the use of particular weapons are among the oldest provisions in the law of war and are inextricably woven into the fabric of modern IHL [50]. While there have been specific conventions restricting or prohibiting particular weapons such as chemical weapons or landmines, IHL also contains general principles such as the principle of distinction and proportionality which apply as a matter of customary international law [19]. For our framework, we will primarily rely the latter for two reasons. First, there evidently is no specialised normative convention as of yet for AI weapons, as the matter is still under discussion before the CCW Conferences in Geneva at the time of writing. Second, even if such a convention would exist, it is highly unlikely that all States would accede to it [51]. It is therefore in any situation relevant to consider more general IHL as a unifying normative standard applicable to all States.

As referenced in the introduction, there is little doubt that IHL applies to AI weapons. While modern IHL was born in the twentieth century and many new technologies have been introduced since then (e.g. precision weapons, cyber-weapons, AI), any new weapon is to conform itself to the applicable rules, and not vice versa [25]. IHL is applicable "without regard to the kind of technology

in question" [17]. This is confirmed consistently throughout the debate on AI weapons. While there is some contention on whether existing IHL is necessarily *sufficient* to regulate all challenges that arise from AI in weapon systems [24], it is uncontroversial that IHL continues to apply for the use of military AI [18]. We echo Canada's [6] position that ensuring the lawfulness of AI weapons should be "constant reference points" for any discussion on the matter.

The ability of weapons to fulfil IHL requirements must be tested as early as the development, testing and adoption stages [23]. For parties to Additional Protocol I (API), this is explicitly provided in the form of a duty to conduct an 'Article 36' legal review [2]. Nevertheless, in our discussion, we will focus more on the operational half of the weapon's lifecycle. The reason for this is that reviews are conducted with respect to the *envisaged* tasks and circumstances of use for that particular weapon, and not *in abstracto* [37]. An example can be drawn from legacy weapons. During the Gulf War, Iraq was broadly condemned for the use of SCUD missiles, which had rudimentary guidance systems (i.e., a low accuracy rate), against Israeli population centres. However, such a weapon might validly pass a legal review if it was designed to be deployed only in civilian-free locations [40]. Similarly, an AI weapon with a very low specificity rate for distinguishing between military and civilian objects is not necessarily indiscriminate if, for instance, it is designed to operate underwater [46]. We argue, for this reason, that it is of particular importance to highlight the deployment phase of an AI weapon, i.e., the law that applies to military operations. Any generalised conclusions that may be drawn for the purposes of legal review (e.g. accuracy rates) can subsequently be drawn from the principles applicable to operations.

When we speak of the *operational* half of a weapon's lifecycle (i.e., deployment and use), the targeting cycle becomes our primary reference point. The aim of this process is to synchronise the choice of weapon, target and operational constraints to obtain the desired military effect [10]. Crucially, this effect must be achieved *while* ensuring compliance with IHL [8]. For this reason, militaries directly integrate legal tests into the targeting process to ensure that any weapon that is being considered meets the standards required for lawful use. Unfortunately, this operational perspective has not garnered the attention it deserves: Ekelhof [15] notes that much of the discussion on AI in weapon systems fails to properly take into account the military targeting process. We will therefore place greater scrutiny on this military art of targeting and the way IHL principles are implemented in practice during concrete operational circumstances.

2.1 The targeting cycle

In this section, we summarise the key steps of the targeting cycle based on the US military and NATO standard. While specific details and protocols may differ between military organisations, there the six steps explained below are generally applicable to most military operations.[4]

[4] The following overview is drawn from [32, 13, 35, 15].

(1) Goal analysis involves the commander analysing the broader goals previously set at the strategic or operational levels. For instance, in the NATO context, goals, target sets and guidance are generally provided by the Joint Force Commander. The commander considers the desired end state described by the broader goals and how to best achieve it.

(2) Target analysis, also called target development, involves the identification and specification of eligible targets. From this analysis, a general calculus is made of the action, time, and resources required to engage this target, to what extent this would contribute toward achieving the goals determined in Step (1), and whether there is a risk of collateral damage in view of its location, function, or characteristics.

(3) Capability analysis involves an assessment of the means and methods of warfare available to the commander [8]. It is during this phase that a weapon or weapons mix is selected which would best achieve the desired effects in light of details of the potential targets identified in Step (2). The art of comparing different alternatives and assigning the optimal combination, quantity and delivery of weapons (systems) to minimise collateral damage to the furthest extent possible while still achieving the desired objective is frequently referred to as *weaponeering* [44].

(4) Capability assignment features the definitive matching of the chosen capability mix to the targets. If necessary, the commander can order additional operational safeguards or considerations to be implemented. The assigned order is then forwarded to component commanders for final planning and execution.

(5) Execution takes place at the tactical level and features the operation being carried out based on the considerations made during all previous phases. A miniature version of the targeting cycle is performed here by the component commander. At some point, a decision to use force is made that cannot be undone, such as the drawing of the trigger on a sniper rifle or the launching of a weapon that cannot be recalled [35]. This is called the *execution moment*.

(6) Assessment is a crucial element in the iterative targeting procedure. Any change in the environment as a result of Step (5) is assessed, noted, and reported back to determine the impact of the use of force both in terms of achieving the desired military effect and damage to civilians. From this new information, the operational goals are re-assessed and the process begins anew in Step (1) until the desired military objective is achieved [22]. Additionally, even if no new engagements are planned, results from Step (6) are still recorded for the purposes of general after-action reviews and lessons learnt, both important processes for preventing the repetition of mistakes in future operations [32].

A graphical representation of this process can be found in Figure 1.

2.2 Integrating the law into the cycle

Certain obligations in IHL are considered to be inextricably linked to the targeting process. These requirements are for a major part to be found in Article 57 of API. For instance, the UK Ministry of Defence [47] notes that "any system, before an attack is made, must verify that targets are military entities, take

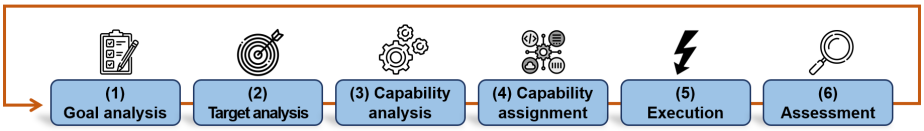

Fig. 1: Graphical illustration of the targeting cycle

all feasible precautions to minimise civilian losses and ensure that attacks do not cause disproportionate incidental losses". In Thurnher's [45] view, targeting requires "examination of three key requirements of the law of armed conflict: distinction, proportionality, and precautions in the attack". A summation of these obligations, however, does not provide us with an indication of when and how they are exactly applied within the operational context. For this reason, we take a closer look at how militaries implement these legal tests when executing the targeting cycle.

As with the individual steps of the targeting cycle itself, there is no universal template to fall back upon, but there is usually an efficient order adopted by most militaries. Some authors, such as Corn [8] and Ducheine & Gill [14], have proposed flowcharts to this effect as shown in Figure 2. Corn's approach is more akin to a decision tree, while Ducheine & Gill's approach better illustrates how individual legal tests are timed within the 6 steps illustrated in Figure 1. The latter also demonstrates well the effects of different variables such as collateral damage and military advantage and that the process can skip tests or loop around, depending on the applicable inputs.

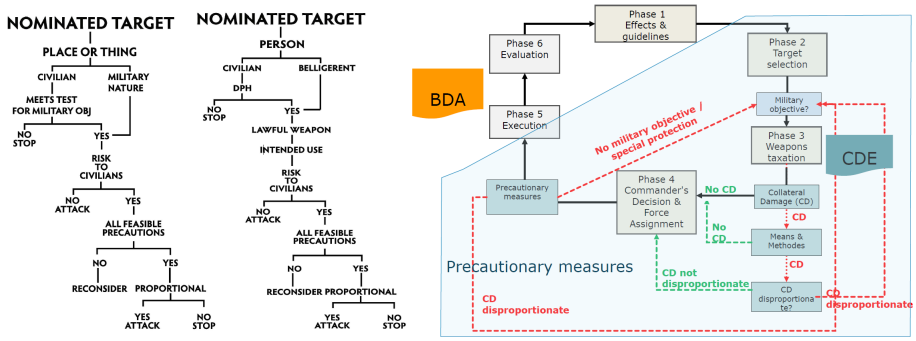

Fig. 2: Proposed flowcharts by Corn [8] and Ducheine & Gill [14]

While such flowcharts are useful for human commanders, for our hypothetical AI system, we need to extract more clearly the specific legal tests, how they are executed, and what inputs are required for each. We expand on this now.

3 Timing and application of legal tests

No particular legal test is applied during **(1) Goal analysis**. However, the goal and rationale of the operation itself may have ramifications in terms of defining the importance of particular objectives or targets, i.e., the military advantage that can be gained. We will thus refer back to these goals in later phases when their legal relevance becomes more apparent.

(2) Target analysis features several important legal requirements. First, the principle of distinction (API Art.48) requires the collection of information and intelligence to ensure that the envisaged target(s) are indeed valid objectives. IHL asks attackers to "verify that the objectives to be attacked are neither civilians nor civilian objects and are not subject to special protection but are military objectives" (API Art.57(2)(a)(i)). Step (2) is usually deemed the ideal moment to apply this test [14]. Militaries also conduct collateral damage estimation at this phase, i.e., whether there is risk of incidental civilian harm tied to the target [32, 44].

A specific form of precautions found in ***API Art.57(3)***, and our first main legal test, can be applied at this stage [35]. This paragraph requires the following: "When a choice is possible between several military objectives for obtaining a similar military advantage, the objective to be selected shall be that the attack on which may be expected to cause the least danger to civilian lives and to civilian objects". This test involves two variables: military advantage and collateral damage, both of which can be derived from Step (1) and Step (2) respectively. While both variables are complex to quantify, the rule itself is relatively straightforward. If by $D = \{D_1, D_2, ...\}$ we denote a set of possible Decisions (*in casu*, attacking a particular target), and we define the Military Advantage gained from decision D_t as MA_t and the collateral damage involved from decision D_t as IH_t (from 'Incidental Harm') in a binary situation between target D_1 and D_2, we could formalise the rule as follows:

$$\forall_{D_x, D_y \in D} \textbf{ if } MA_x \approx MA_y \wedge IH_x < IH_y \textbf{ then } \text{select } D_x \textbf{ else } \text{select } D_y \quad (1)$$

(3) Capability analysis is the most involved step in terms of legal tests. The two most important principles addressed at this stage concern proportionality and minimisation, which both relate to collateral damage. First, ***proportionality*** prohibits any Decision (i.e., a combination of a target, capability and method of delivery) which causes incidental civilian harm excessively disproportionate to the concrete military advantage anticipated (API Art.57(2)(a)(iii)). As with the test in Art.57(3) above, this involves a comparison between military advantage and collateral damage. Multiple options can be considered 'proportionate' as long as the threshold of excessiveness is not exceeded [31]; in other words, proportionality effectively sets a maximum threshold of how much collateral damage remains acceptable vis-à-vis the anticipated military advantage. If we define this threshold as p, we can formalise[5] the rule as:

[5] A similar approach to the modeling of proportionality rule can be found in [52].

$$\textbf{if for } D_x \text{ s.t. } \frac{IH_x}{MA_x} \leq p \textbf{ then } status(D_x) = legal \textbf{ else } status(D_x) = illegal \tag{2}$$

It is not sufficient to only look at proportionality. IHL also mandates that incidental harm to civilians must be ***minimised*** to the furthest extent feasible (API Art.57(2)(a)(ii)). Essentially, "[i]f there is a choice of weapons or methods of attack available, a commander should select those which are most likely to avoid, or at least minimize, incidental damage" [26]. This obligation involves comparing different options in terms of capabilities, operational constraints and methods of delivery. For instance, applying this test may result in less accurate weapons being discarded, altering the timing of attack, or selecting a less destructive damage mechanism [37]. Commanders are not expected to do the impossible: the corollary of *feasibility* empowers commanders to take into consideration all relevant circumstances, including those relevant to the success of military operations [37]. If too much military advantage is lost due to a particular minimisation measure, they are permitted to select a more reasonable option that better balances the humanitarian and military considerations in play [40]. In a binary comparison therefore, this rule would be formulated:

$$\textbf{if } \exists_{D_y \in D} \forall_{D_x \in D \setminus D_y} : \frac{IH_y}{MA_y} < \frac{IH_x}{MA_x} \textbf{ then select } D_y \tag{3}$$

It must also be emphasised that because proportionality and minimisation are both concerned with the variable of collateral damage (IH), the tests can be disregarded in cases where this variable does not factor [14]. If the weapon is projected to function only in a military-exclusive environment, for example, these tests may be skipped.

Finally, there is one additional requirement that must be mentioned at this stage related to weapons which are ***inherently illegal***. Customary law generally recognises two aspects which make a weapon unlawful per se: that of causing unnecessary suffering and of being inherently indiscriminate [37, 45]. In addition, ***weapons treaties*** may be in force for the belligerent which limit the use of certain weapons beyond what customary law requires [4]. While these are important legal restrictions, it is irregular for these weapons to reach the targeting stage: It is the role of weapons reviewers to filter out such weapons during development and adoption [23, 48]. Nevertheless, it cannot be ruled out that a weapon fulfilling such criteria actually reaches the front lines at some point. To guard against this possibility and maintain IHL-compliance, we add this function to allow the AI to deny its own use if its deployment would be unnecessarily injurious, inherently indiscriminate or prohibited by treaty.

During **(4) Capability assignment** up to the **execution moment**, a series of final precautionary measures are enacted. In part these involve continuous ***re-tests*** of all previous obligations on a more detailed level [35]. The reason for this is to ensure that all input assumptions related to the classification of the target, its military worth and the collateral damage estimations underlying the

previous decisions remain applicable. Related to this, IHL also requires ***cancelling*** the ordered attack if it becomes known throughout this period that any of these assumptions are no longer valid (API Art.57(2)(b)). In addition, advance ***warning*** must be provided in cases where civilians can be affected (API Art.57(2)(c)), although this duty only applies in cases where this is reasonable, i.e., where it would not compromise the success of the attack [37].

Finally, during **(6) Assessment**, a legal obligation that could be relevant is the duty to ***suppress and repress***. It is part of the broader obligation to respect and ensure respect for IHL at all times [1], and involves both general measures to prevent and address violations of IHL ('suppression') and, in case a serious breach has occurred, that the persons responsible for such violations be held criminally responsible ('repression') [1]. Thus, requirements related to foreseeability, understandability, traceability and the keeping of digital records may become relevant at this stage [27].

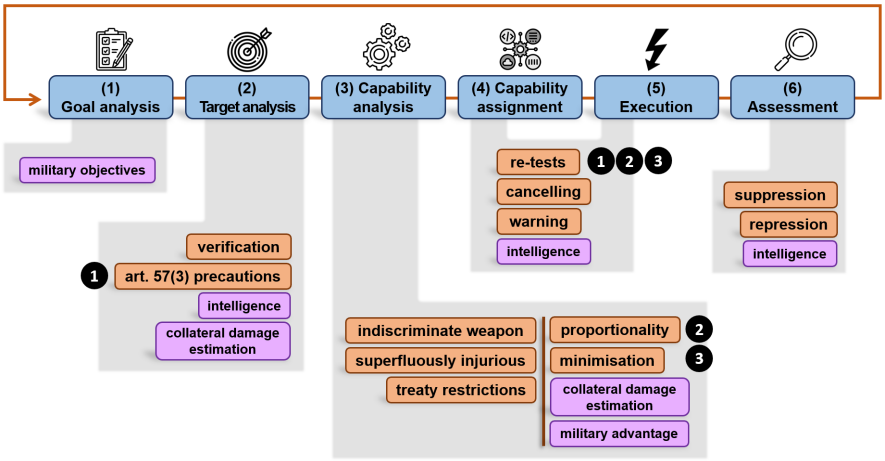

Fig. 3: Overview of legal tests and inputs during the targeting cycle

A graphical summary of targeting phases and the corresponding legal tests, in addition to supporting actions that contribute toward the fulfilment of these legal obligations, are depicted in Figure 3. Legal tests are represented by orange boxes, important inputs by purple boxes, and the formulae described above by black circles. The framework is intended to be function-agnostic, but whether each test is to be applied by a human or an AI will depend on the specific AI under consideration. If the execution of a particular phase is left to an AI, then it must be demonstrated that the AI in question is capable of executing the legal tests necessary for that phase [4].

4 The general framework of the autonomous targeting system

In the previous section, we described the general targeting sequence and the legal tests that must be applied during each step for lawful use of force. In this section, we are going to present the general structure of our autonomous targeting system which incorporates these legal tests into its functionality. Note that we will not discuss any particular targeting scenario, but rather introduce a general (function agnostic) framework which is IHL-compliant. Moreover, we will not discuss the technical details of the machinery which can be used to implement the targeting system, but only propose a structure and possible techniques which may allow for creation of a targeting system which can observe the legal requirements identified in Section 2. Further technical details of this system will be discussed in future papers. We realize that some functionalities may still be very difficult to implement in real life systems (e.g. distinguishing between combatants and civilians) [43]. However, we can expect that such modules, at least for some tasks (e.g. distinguishing military and civilian aircraft), will be feasible in the near future.

One of the most important assumptions on the basis of which we designed our model is rooted in the observation that although the transparency and explainability requirements are crucial for many legal tests [27], the requirements for the cognitive elements of the decision process are less restrictive. IHL is a purely normative framework and does not provide rigid requirements or standards for commanders when conducting tests which involve MA (military advantage) and IH (incidental harm) as variables. In fact, the opposite is true: these assessments are frequently described as eminently qualitative and subjective in nature [3]. On the basis of the above, we can assume that at least some functions of the cognitive part can be created with the use of much less transparent data-driven approaches, especially deep learning neural networks.

One prominent element of the procedure described in the previous section is the comparison between anticipated military advantage and anticipated incidental harm. Obviously, while making his decision, a human commander does not represent either variable in a quantifiable form. An autonomous AI-driven model, however, requires not only a quantifiable representation, but also a representation which allows for their formal comparison [5].

We will use values as a central concept allowing for representation of both military advantage and incidental harm. Values we understand as an abstract (trans-situational) concept which allows for the estimation of a particular action or a state of affairs and which influences one's behavior. Consequently, on the basis of such a definition, we assume that particular values can be satisfied to a certain degree [53]. Such a definition of value can be seen as a kind of abstraction of concrete results of an action, and allows us to use them as a central concept in our model where they play an important role as an intermediate concept representing an abstraction of a targeting situation.

On the basis of the above, we provide a discussion of how, from a technical viewpoint, the requirements of each stage of the targeting process can be fulfilled.

Since this is a position paper presenting the overall structure of the system, we will not enter into the technical details of particular functions used in the model, unless it is necessary to make the model understandable or when it constitutes a key element of the discussion. More detailed presentations of the technical nuances will be included in future work.

A battlefield is a multiagent environment par excellence featuring many allied, neutral and adversarial agents which the system must be able to understand and account for [36]. In our framework, we present all functions as if they are fulfilled by a single agent for simplicity. In actual systems, it is possible that several agents are involved which contribute together toward the execution of the framework's different functions, or that particular functions are performed by distinct agents working together. Additionally, input data such as signal intelligence may be obtained from an allied agent, which can either be a human observer or another AI unit such as a reconnaissance drone [49]; similarly, the final decision resulting from our framework can be effectuated by an agent on the frontlines such as a human squad or combat robot. These permutations do not affect the viability of our framework as long as all functions are executed correctly by the agents involved and, in the case of collaborating agents, all necessary tests take into consideration all involved agents. With regard to adversarial agents, sufficient robustness and adaptability against opponents' efforts to disrupt the system's proper functioning must be made into important design requirements [12].

5 The structure of the system

In this section we introduce the general structure of the proposed system.

5.1 The basics of the model

Firstly, we discuss the basics:

- **Introduction of goals**. In the first stage of the targeting process, the commander performs the analysis of the desired state on the strategic and operational levels. Since such an analysis is performed from a broader perspective, taking into account the general goals of military operations, we argue that for the autonomous device, such a goal can be represented as a set of thresholds of a group of values which constitute a more general value *military advantage*.
- **Input data**. In order to perform all required tests and to decide which decision should be made, some necessary data has to be prepared. Firstly, the agent should distinguish a set of available actions with their anticipated results and evaluate them in the light of MA and IH. In order to fulfill this stage, a set of preparatory tasks should be performed:
  - Generation, on the basis of signal intelligence and the general circumstances of the case (denoted by S), of the set of decisions which can possibly be made in given circumstances. By D we denote a set of available decisions.

- Prediction of the result of every decision from the set. Note that for the tests described in the previous section, the levels of MA and IH relate to the *anticipated* results of decisions, which means that they are by nature uncertain. Let R be a set of all possible results of actions (decisions from set D) and let PR be a set of conditional probabilities of those results, given a particular situation and decision.
- Evaluation of the decision results in the light of the set of relevant values. Suppose a set of decision results R and a set of functions Φ_V which returns the level of satisfaction of a particular value v_x by result r_y. By VR we denote a set of levels of satisfaction of all values by the results of all available decisions.
  Since function Φ has a crucial character for our model, we briefly present here how it can be obtained. The goal is to find a function which for every possible result ($r \in R$) can predict the level of satisfaction of every value (the level to which a predicted result of a decision would satisfy the relevant value, e.g. *military advantage*, *life of civilians*, etc.). Suppose that every result from set R (possible results of actions) will be evaluated and labelled by human annotators in the light of every value (by assigning a number representing the level of satisfaction of a given value). On the basis of such data and a ML-based regression mechanism, a regression function can be trained which can predict the level of satisfaction of a given value on the basis of a particular result. A more detailed analysis and discussion of this approach will be presented in future work.
- Calculation of the expected level of satisfaction of a particular value (it can be calculated on the basis of probabilities of results PR and evaluation of the decisions' results VR.) By EV we denote a set of levels of satisfaction of all values by all available decisions.

5.2 The structure

In this section we present the structure of the proposed system:

- *Extraction of available decisions* is responsible for obtaining a set of available decisions (S is an input, D is an output of the module).
- *Result prediction module* is responsible for predicting results of decisions with their probabilities (D and S are inputs to the module, while PR and R are outputs).
- *Evaluation module* is responsible for performing function ϕ (G, S, and R are inputs to the module, VR is an output).
- *Parameters' extraction module* is the module which returns the set of parameters of decisions. By the parameters of a decision we understand details of a decision such as type of weapon, timing, etc. D is an input, PAR is an output of the module.
- *Expected evaluation module* is responsible for calculating the expected evaluation of decisions in the light of values (VR and PR are an input, EV is an output of the module).

- *Treaties fulfillment module* is responsible for performing function filtering decisions which do not fulfill treaties (PAR and $FPAR$ are inputs, DTR is an output of the module). If by $FPAR$ we denote the set of requirements imposed by treaties, then the module can work as a logic-based reasoning mechanism.
- *Goals fulfillment module* is responsible for performing function filtering decisions which do not fulfill the commander's goals (G and EV is an input, DG is an output of the module). If by a goal we understand the minimal acceptable levels of values' satisfaction (see [53]), then a given decision will fulfill the goal if the expected level of satisfaction of relevant values will be above the thresholds assumed in G.
- *Harm minimization filter* is responsible for the process of minimization of incidental harm (EV is an input, DMH is an output). A given decision will pass the test if for this decision formula 3 will be fulfilled.
- *Proportionality test* is responsible for performing the proportionality test (EV is an input, DP is an output). A given decision will pass the test if for this decision formula 2 will be fulfilled.
- *Article 57(3) Filter* is responsible for the process of filtering decisions which for the same military advantage causes greater harm to civilians (Article 57(3), EV is an input, DT is an output). A given decision will pass the test if for this decision formula 1 will be fulfilled.
- *Fulfillment of requirements* is responsible for joining together results of the above tests (DT, DP, DMH, DG, and DTR are inputs and DAV is an output). A given decision will fulfill this requirement if all tests have been passed.
- *Decisions ordering* is responsible for ordering available decisions (those fulfilling the above tests) on the basis of the level of satisfaction of Military Advantage (DAV and VR is an input, *Decisions* is an output of the module).

The structure of the proposed model is presented in Figure 4. The model features a clear distinction between (1) the cognitive part of the decision process, including functions extracting available decisions, their results, and evaluation (the upper part of the scheme) and (2) the reasoning part of the decision process, including legal tests, goal test, treaties test, etc. (lower part of the scheme). This distinction between the parts of the decision process is a notable strength of the framework we propose because it provides some degree of transparency and explainability. These attributes have been identified as crucial both for the lawful use of AI weapon systems and upholding the responsibility of its users [27].

As such, the structure we propose relies on the conviction that for the sake of transparency, legal tests should be performed in an explainable way, i.e. the system should explicitly check whether a given decision passes all necessary tests, while the other elements of the decision process can utilize data-driven approaches. Such an approach is compliant with the general approach regarding hybrid systems, in which the data-driven part is used for extraction of the input

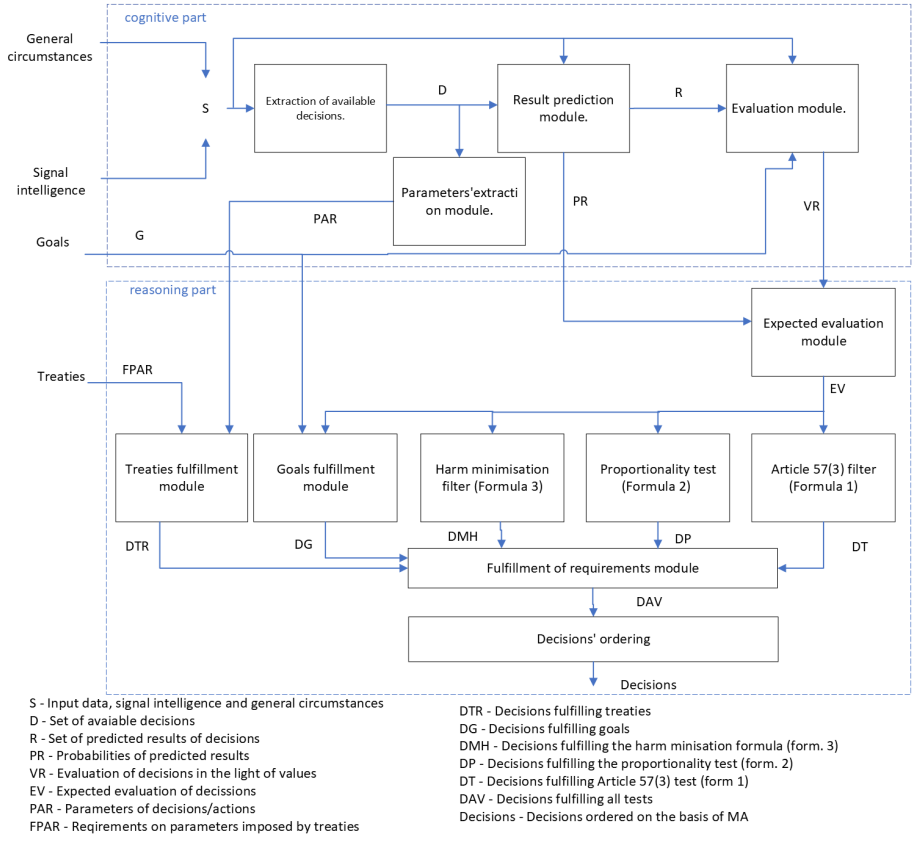

Fig. 4: Graphical illustration of the system's structure

data for a knowledge-based system, and generally allows for filling the so-called semantic gap between data and knowledge [29].

6 Discussion and Conclusions

The paper introduces a framework for creating an AI-based hybrid targeting system for military autonomous agents capable of operating within the bounds of IHL. The main goal was to present a way how IHL can be integrated from the ground up into a military AI system in order to better guarantee IHL-compliance. We present the main stages of the targeting process, identify which legal requirements are imposed by IHL and what variables and elements these tests encompass, and introduce a mechanism which allows for the development of a system fulfilling those requirements.

To achieve this, we introduced a model of a hybrid system which combines data-driven parts (possibly created with the use of deep learning neural net-

works) and knowledge-driven parts. This type of system composition allows us to draw from the advantages of both AI paradigms, while also compensating for at least part of their respective disadvantages. In particular, one major disadvantage of data-driven AI, lack of transparency, is overcome to some extent, which is a boon for IHL compliance.

Further development of our framework requires the verification of the model. Since this paper presents a general model of the decision-making process only, we cannot introduce here a fully-fledged, technical verification of our proposal. Instead we briefly sketch how the verification of the model can be performed. Since our framework consists of two parts - cognitive and reasoning ones - the verification should be performed twofold:

- The cognitive part should be verified on the basis of statistical quality of all modules. For example, the quality of sensors, the accuracy of predictions and evaluations, etc. The verification of this part is task-dependent. For example, signal intelligence should be verified in the light of accuracy of object detection related to the specific sensors: cameras, recorded sound, satellite pictures etc.; the prediction module should be evaluated in the light of accuracy of predictions made; etc. Every module should be verified in the context of the concrete intended purpose for which the device is designed [28].
- The reasoning part requires formal and legal verification of all tests (see section 3 where we discuss some legal aspects concerning the model) and the whole reasoning process and the formal and experimental analysis of the reasoning machinery used to performing necessary tests, which will be presented in our future works.

Our hypothetical framework was developed with the aim of identifying and elaborating the functionalities which would be necessary for AI-driven systems to conform to IHL. We make no practical pronouncements concerning technical implementation or in what type of weapon this framework would be incorporated, as these details would depend on the military organisation's specific needs. In addition, it is possible that comparable systems are currently under development by militaries. These systems are likely to remain confidential and thus, it is difficult for us to test our framework vis-à-vis those systems. Our proposal nevertheless can be used as a reference or guideline for both current and future constructors intending to build systems with IHL compliance in mind.

References

1. Geneva Convention for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field (adopted 12 August 1949, entered into force 21 October 1950) 75 UNTS 31
2. Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (adopted 8 June 1977, entered into force 7 December 1978) 1125 UNTS 3 (1977)

3. Anderson, K., Waxman, M.C.: Law and Ethics for Autonomous Weapon Systems: Why a Ban Won't Work and How the Laws of War Can (2013)
4. Boothby, W.H.: Regulating New Weapon Technologies. In: Boothby, W.H. (ed.) New Technologies and the Law of War and Peace, pp. 16–42. Cambridge University Press, Cambridge s (2019)
5. Boulanin, V.: Mapping the development of autonomy in weapon systems: A primer on autonomy. Stockholm International Peace Research Institute, Stockholm (2016)
6. Canada: Opening Statement by Canada at Second Meeting of the Group of Governmental Experts on Lethal Autonomous Weapons Systems (LAWS), 9–13 April 2018. Tech. rep. (2018)
7. Chavannes, E., Arkhipov-Goyal, A.: Towards Responsible Autonomy: The Ethics of Robotic and Autonomous Systems in a Military Context. The Hague Centre for Strategic Studies, The Hague (2019)
8. Corn, G.S.: War, law, and the oft overlooked value of process as a precautionary measure. Pepperdine Law Review **42**, 419–466 (2014)
9. Crootof, R.: The Killer Robots are here: Legal and Policy Implications. Cardozo Law Review **36**, 1837–1915 (2015)
10. Curtis E. Lemay Center: Air Force Doctrine Publication 3-60 - Targeting (2019), www.doctrine.af.mil/Doctrine-Publications/AFDP-3-60-Targeting
11. Dahlmann, A., Dickow, M.: Preventive Regulation of Autonomous Weapon Systems. Tech. Rep. Stiftung Wissenschaft und Politik Research Paper 2019/RP 03, Berlin (2019). https://doi.org/10.18449/2019RP03
12. Defense Innovation Board: AI Principles : Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense Defense Innovation Board. Tech. rep., Department of Defense (2019), https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF
13. Department of the Army: The Operations Process (2019)
14. Ducheine, P., Gill, T.: From Cyber Operations to Effects: Some Targeting Issues. Militair Rechtelijk Tijdschrift **111**(3), 37–41 (2018)
15. Ekelhof, M.: Human control in the targeting process. In: Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons, pp. 53–56. ICRC, Versoix (2016)
16. Eklund, A.M.: Meaningful Human Control of Autonomous Weapon Systems: Definitions and Key Elements in the Light of International Humanitarian Law and International Human Rights Law. Totalförsvarets forskningsinstitut, Stockholm (2020)
17. Geiß, R., Lahmann, H.: Autonomous weapons systems: a paradigm shift for the law of armed conflict? In: Ohlin, J.D. (ed.) Research Handbook on Remote Warfare, pp. 371–404. Edward Elgar, Cheltenham (2017)
18. Group of Governmental Experts on Lethal Autonomous Weapons Systems (GGE on LAWS): Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, UN document CCW/GGE.1/2019/3, 25 September 2019. Tech. rep., Geneva (2019)
19. Henckaerts, J.M., Doswald-Beck, L.: Customary International Humanitarian Law, Volume I – Rules. ICRC, Geneva (2005)
20. Holland Michel, A.: The Black Box, Unlocked: Predictability and Understandability in Military AI. Tech. rep., United Nations Institute for Disarmament Research, Geneva, Switzerland (sep 2020). https://doi.org/10.37559/SecTec/20/AI1, https://unidir.org/black-box-unlocked

21. House of Lords: Select Committee on Artificial Intelligence, Report of Session 2017-19, AI in the UK: Ready, Willing, and Able? Tech. Rep. HL Paper 100, 16 April 2018 (2018)
22. Huffman, W.B.: Margin of Error: Potential Pitfalls of the Ruling in The Prosecutor v. Ante Gotovina. Military Law Review **211**, 1–56 (2012), hdl.handle.net/10601/2104
23. International Committee of the Red Cross: A Guide to the Legal Review of New Weapons, Means and Methods of Warfare: Measures to Implement Article 36 of Additional Protocol I of 1977. ICRC, Geneva (2006)
24. International Committee of the Red Cross: 'Report of the ICRC Expert Meeting on 'Autonomous weapon systems: technical, military, legal and humanitarian aspects', 26-28 March 2014, Geneva'. Tech. rep. (2014)
25. International Court of Justice: Legality of the Threat or Use of Nuclear Weapons (1996)
26. International Criminal Tribunal for the Former Yugoslavia: Final Report to the Prosecutor by the Committee Established to Review the NATO Bombing Campaign Against the Federal Republic of Yugoslavia. Tech. rep. (2001), www.icty.org/en/press/final-report-prosecutor-committee-established-review-nato-bombing-campaign-against-federal
27. Kwik, J., Van Engers, T.: Algorithmic fog of war: When lack of transparency violates the law of armed conflict. Journal of Future Robot Life pp. 1–24 (jul 2021). https://doi.org/10.3233/FRL-200019
28. Meier, M.W.: Lethal Autonomous Weapons Systems. In: Complex Battlespaces, pp. 289–316. Oxford University Press, Oxford (jan 2019). https://doi.org/10.1093/oso/9780190915360.003.0010
29. Meyer-Vitali, A., Bakker, R., van Bekkum, M., de Boer, M., Burghouts, G., van Diggelen, J., Dijk, J., Grappiolo, C., de Greeff, J., Huizing, A., Raaijmakers, S.: Hybrid ai white paper. Tech. rep., TNO (2019), tNO 2019 R11941
30. Ministère des Armées (France): L'intelligence artificielle au service de la défense. Tech. rep., Ministère des Armées, Paris (2019)
31. Neuman, N.: Applying the rule of proportionality: force protection and cumulative assessment in international law and morality. Yearbook of International Humanitarian Law **7**, 79–112 (dec 2004). https://doi.org/10.1017/S1389135904000790
32. North Atlantic Treaty Organisation: Allied Joint Doctrine for Joint Targeting, Edition A Version 1 (April 2016) AJP-3.9 (2016)
33. Office of the Assistant Secretary of Defense for Research and Engineering: Technical Assessment: Autonomy. US Department of Defense, Washington, D.C. (2015)
34. Roff, H.M.: Meaningful Human Control or Appropriate Human Judgment? The Necessary Limits on Autonomous Weapons (2016)
35. Roorda, M.: NATO's Targeting Process: Ensuring Human Control Over (and Lawful Use of) 'Autonomous' Weapons'. In: Williams, A.P., Scharre, P.D. (eds.) Autonomous Systems: Issues for Defence Policymakers, pp. 152–168. NATO, The Hague (2015)
36. Russell, S.J., Norvig, P.: Artificial Intelligence: A Modern Approach. Pearson, New Jersey, 3 edn. (2010)
37. Sandoz, Y., Swinarski, C., Zimmerman, B.: Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949. Martinus Nijhoff (1987)
38. Scharre, P., Horowitz, M.C.: An Introduction to Autonomy in Weapon Systems. Tech. rep., Center for a New American Security (2015)

39. Scharre, P.D.: The Opportunity and Challenge of Autonomous Systems. In: Williams, A.P., Scharre, P.D. (eds.) Autonomous Systems: Issues for Defence Policymakers, pp. 3–26. NATO, The Hague (2015)
40. Schmitt, M.N., Garraway, C.H., Dinstein, Y.: The Manual on the Law of Non-international Armed Conflict, With Commentary. International Institute of Humanitarian Law, San Remo (2006)
41. Schuller, A.: At the Crossroads of Control: The Intersection of Artificial Intelligence in Autonomous Weapon Systems with International Humanitarian Law. Harvard National Security Journal **8**, 379 (2017)
42. Sharkey, N.E.: Towards a Principle for the Human Supervisory Control of Robot Weapons. Politica & Società p. 305 (2014)
43. Szpak, A.: Legality of Use and Challenges of New Technologies in Warfare – the Use of Autonomous Weapons in Contemporary or Future Wars. European Review **28**(1), 118–131 (feb 2020). https://doi.org/10.1017/S1062798719000310
44. Thorne, J.G.: Warriors and War Algorithms: Leveraging Artificial Intelligence to Enable Ethical Targeting. Tech. rep., Naval War College (2020), https://apps.dtic.mil/sti/citations/AD1104171
45. Thurnher, J.S.: Examining Autonomous Weapon Systems from a Law of Armed Conflict Perspective. In: Nasu, H., McLaughlin, R. (eds.) New Technologies and the Law of Armed Conflict, pp. 213–228. T.M.C. Asser Press, The Hague (2014)
46. Thurnher, J.S.: Feasible Precautions in Attack and Autonomous Weapons. In: von Heinegg, W.H., Frau, R., Singer, T. (eds.) Dehumanization of Warfare: Legal Implications of New Weapon Technologies, pp. 99–117. Springer International Publishing AG, New York (2018)
47. UK Ministry of Defence: The UK Approach to Unmanned Aircraft Systems: Joint Doctrine Note 2/11. Tech. rep., United Kingdom Ministry of Defence (2011)
48. United States Office of General Counsel of the Department of Defense: Law of War Manual, Updated December 2016. Tech. rep., Department of Defense (2015)
49. U.S. Air Force Office of the Chief Scientist: Autonomous Horizons: System Autonomy in the Air Force— A Path to the Future, Volume I: Human-Autonomy Teaming. Tech. Rep. AF/ST TR 15-01 (2015)
50. Wallace, D.: Cyber Weapon Reviews under International Humanitarian Law: A Critical Analysis, Tallinn Paper no 11. Tech. rep. (2018)
51. Wilson, C.: Artificial Intelligence and Warfare. In: Martellini, M., Trapp, R. (eds.) 21st Century Prometheus Managing CBRN Safety and Security Affected by Cutting-Edge Technologies, pp. 141–177. Springer Nature Switzerland AG, Cham (2020)
52. Zurek, T., Woodcock, T., Pacholska, M., van Engers, T.: Computational modelling of the proportionality analysis under international humanitarian law for military decision-support systems. https://ssrn.com/abstract=4008946 (January 2022)
53. Zurek, T.: Goals, values, and reasoning. Expert Systems with Applications **71**, 442 – 456 (2017). https://doi.org/http://dx.doi.org/10.1016/j.eswa.2016.11.008