February 2023

ELSA LAB
DEFENCE

No. 2

Research paper series

Model of a military autonomous device following International Humanitarian Law

AUTHORS

Tomasz Zurek, Jonathan Kwik, Tom van Engers

www.elsalabdefence.nl

Ethics and Information Technology (2023) 25:15
https://doi.org/10.1007/s10676-023-09682-1

ORIGINAL PAPER

Model of a military autonomous device following International Humanitarian Law

Tomasz Zurek[1] · **Jonathan Kwik**[2] · **Tom van Engers**[3]

Published online: 15 February 2023
© The Author(s) 2023

Abstract
In this paper we introduce a computational control framework that can keep AI-driven military autonomous devices operating within the boundaries set by applicable rules of International Humanitarian Law (IHL) related to targeting. We discuss the necessary legal tests and variables, and introduce the structure of a hypothetical IHL-compliant targeting system.

Keywords Autonomous military system · International Humanitarian Law · Decision making

Introduction

Many States have expressed interest in incorporating more AI into their military systems, particularly to replace certain aspects of human decision-making during operations (Defense Science Board, 2012; Ministère, 2019). A variety of tasks are envisaged including decisions concerning the use of force, which involve functions such as target selection, sorting, prioritisation and engagement (Thurnher, 2018; Abaimov & Martellini, 2020). While such developments are welcomed by some, others resist the idea of delegating use of force decisions to AI for a variety of reasons, such as difficulties in assigning responsibility (Chengeta, 2016; Bo, 2021), fears that it might reduce the threshold for conflict (Hitoshi & McLaughlin, 2014; Crootof, 2015), and ethical concerns such as dehumanisation (International Committee of the Red Cross, 2018; Sartor & Omicini, 2016) (for summaries of the overall debate and points of contention, see (Lewis, 2015; Cummings, 2018; Santoni de Sio & van den Hoven, 2018; Eklund, 2020)). While the concept of meaningful human control has been interpreted by some to require a human in the loop at all times, others have interpreted the requirement more loosely as also being attainable through meaningful human involvement in the decision-making process leading up to the point of activation (Kwik, 2022). Assuming no consensus can be reached on this matter, it is not precluded that, at least by adherents of the latter school, certain decisions will indeed be delegated to AI in the future. If this is the case, it is important to explore to what extent AI-controlled devices can operate within boundaries set by law. In this paper, we will particularly focus on one frequently-raised (Lewis, 2015; Kalmanovitz, 2016) legal concern: the question of whether such systems can properly comply with international humanitarian law (IHL), as it is often argued (Crootof, 2015; Szpak, 2020; McDougall, 2019) that incorporating many targeting principles such as distinction, proportionality and precautions into an AI is impossible.

In light of this question, in this paper, we advance one way to embed legal rules into military AI by proposing a framework which was built from the ground up on a foundation of IHL principles. The framework is intended to incorporate both IHL rules and to be conscious of practical operational reality, i.e., it is constructed with both the law and military practice in mind. To achieve this, we base our framework on the military targeting cycle, i.e. the set of steps and assessments commanders apply before assigning force. The targeting cycle takes both operational and

✉ Tomasz Zurek
t.zurek@asser.nl

Jonathan Kwik
h.c.j.kwik@uva.nl

Tom van Engers
T.M.vanEngers@uva.nl

[1] T.M.C.Asser Institute, Amsterdam University, R.J. Schimmelpenningcklaan 20-22, The Hague 2517JN, The Netherlands

[2] Faculty of Law, Amsterdam University, Amsterdam, The Netherlands

[3] Complex Cyber Infrastructure, Informatics Institute, Amsterdam University, Amsterdam, The Netherlands

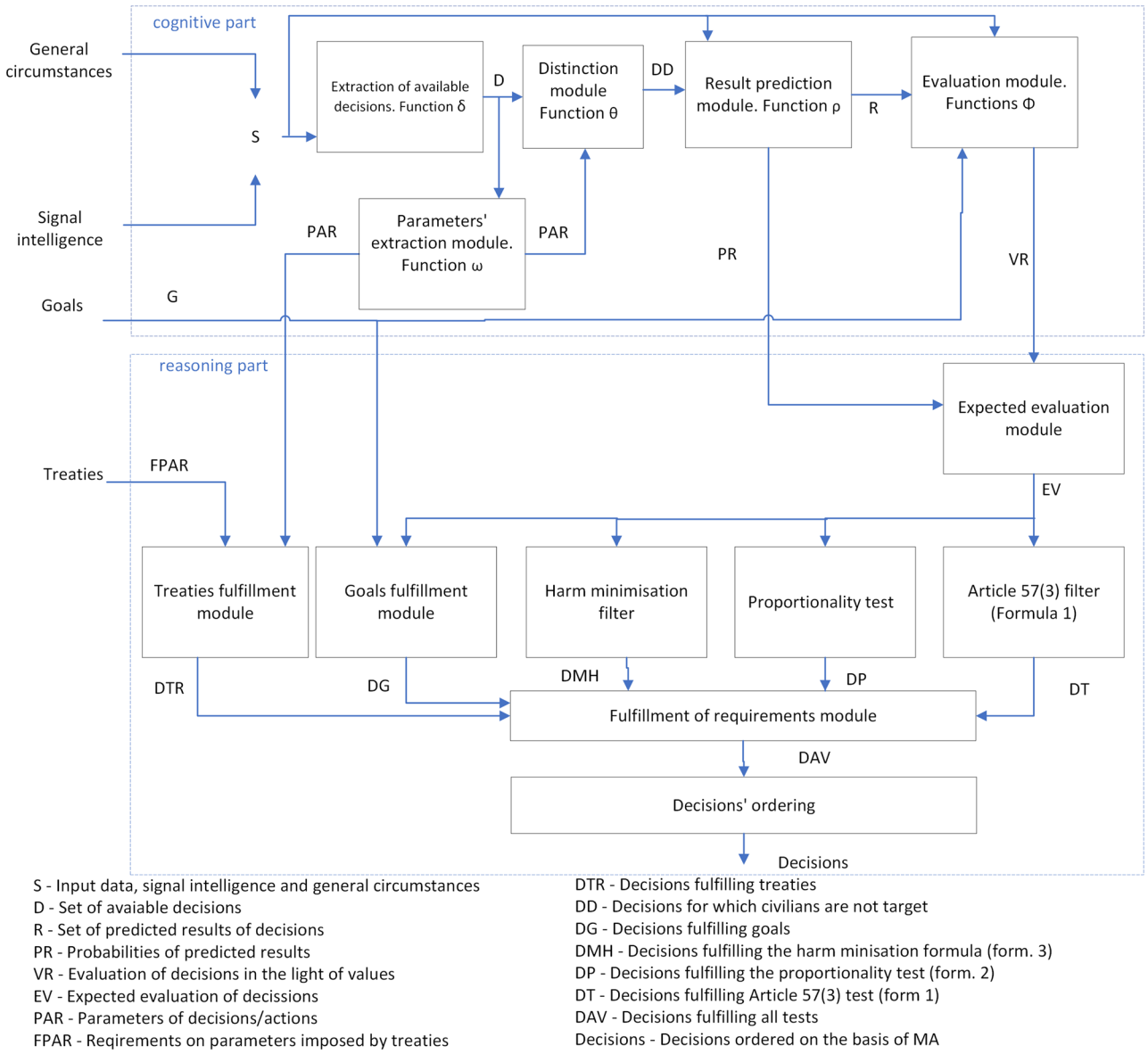

Fig. 1 Graphical illustration of the system's structure

legal considerations as inputs: the former includes variables such as goals, desired effects and military advantage, while the latter includes humanitarian variables such as incidental harm (collateral damage). By basing our system on the targeting cycle, we ensure both practical usability and legal compliance whenever this hypothetical AI is deployed in place of a human-controlled system. We assume that ethical considerations determining under what circumstances these systems (or more generally the use of violence) should be promoted or rejected are reflected in the legal rules set. Our contribution specifically aims to address the legal arm of the discussion.

The principal goal of our work is to discuss the feasibility of implementing autonomous military devices that can follow IHL. We fulfil this goal by distinguishing the key components necessary to perform IHL compliance tests with all the data necessary to perform those tests, and discuss potential problems and threats. In our view, this discussion is especially important because of the specific character of IHL rules incorporating both legal and moral reasoning. Beyond this, the model is agnostic as to which legal tests a deploying commander wishes to delegate to the system. In our model, each legal test is performed by a dedicated module with clear inputs and outputs (see Fig. 1), which provides flexibility. For instance, a commander may wish to perform all other tests themselves, but delegate the harm minimisation function to an autonomous drone, in which case the harm minimisation module would be relevant for

consideration. Alternatively, a commander may be employing a proportionality adviser, in which case we focus on the proportionality module.

Ultimately however, there is broad consensus (Scharre, 2018; Boddens Hosang, 2021; Davison, 2017) that no amount of delegation removes the commander's *legal* responsibility for any consequences of deploying such devices. As Cherry & Johnson (2020) submits, the "legal obligation to take precautions does not fall to the autonomous system". The commander remains responsible for ensuring the attack is lawful and, where they delegate some tasks to autonomous devices, to ensure that required legal tests are performed correctly (Boothby, 2019). In case where a task is indeed delegated, a common conceptualisation is that the human decision-making process is not eliminated but instead moved 'forward' (Ekelhof, 2016; Adviesraad Internationale Vraagstukken, 2015). As mentioned above, the popular concept of meaningful human control as summarised by Eklund (2020); Kwik (2022) is in this regard helpful: even if commanders delegate a task(s) to a device, they remain responsible for ensuring sufficient knowledge of the system's reliability, predicting its approximate behaviour in the field, its robustness against likely enemy countermeasures (International Committee of the Red Cross, 2016), and implementing sufficient contextual controls to ensure performance and controllability is maintained. The main contribution of this work, then, is to provide a general framework of how an autonomous device can perform each of these functions. The casuistic decision of whether these tasks *should* be delegated remains a question of military judgment (Organisation, 2019; van den Boogaard & Roorda, 2021).

We proceed as follows. First, we briefly discuss the foundational framework of our system, the targeting cycle. We also extract and emphasise the most important legal tests that are conducted during this cycle, and formalise them for use in our proposed system. On this basis, we then introduce the building blocks of our system, and discuss its necessary functionalities, its structure and required data. We also provide a brief discussion of the methods used to obtain the data necessary to perform the legal analyses required by IHL.

Targeting and legal tests

As we set out to answer whether targeting rules can theoretically be fulfilled through an AI system, the setup of our model is based on the targeting cycle, i.e. the process taken by commanders that precedes military operations or engagements. It involves analysing potential targets and considering different weapon and delivery options to obtain the optimal military benefit, *while* ensuring compliance with IHL (Corn, 2014). Targeting has important IHL tests already pre-installed into the process, while remaining a preeminently military exercise that is ultimately aimed at achieving the military organisation's objectives and end states in the most efficient way possible (Curtis, 2019). This makes it an ideal foundation for our system.

We analysed the NATO Standard Targeting Procedure (North Atlantic Treaty Organisation, 2016; Roorda, 2015; Ekelhof, 2016) and subsequently extracted the exact legal tests performed and their timing within the cycle, distilling them into six overall steps as follows:

- **(1) Goal analysis** involves the commander analysing the broader goals and objectives previously set at the strategic or operational levels. No legal tests are applied at this juncture, but we observe that these *goals and objectives* constitute an important part of the input.
- **(2) Target analysis** involves the identification and specification of eligible targets. Two important variables come into play here: MA (*Military Advantage* - how much advantage is gained from attacking that target) and IH (*Incidental Harm* - collateral damage caused by both foreseeable direct and indirect effects of the attack). ***Distinction*** and ***Art. 57(3) Precautions*** are applied at this stage.
- **(3) Capability analysis** involves an assessment of what decisions are available (Corn, 2014). For clarity, "*decision*" in our framework refers to any relevant combination of target, weapon system, ammunition, delivery method, etc. Once again, MA and IH are relevant inputs, and are refined further based on the specific weapons mix being envisaged. ***Proportionality***, ***Minimisation*** and ***Discarding Illegal Weapons*** is applied at this stage.
- During **(4) Capability assignment** and up to **(5) Execution**, continuous ***re-tests*** on a more detailed (tactical) level are applied of the previous rules, particularly proportionality, distinction and precautions (Roorda, 2015). If any changes take place which invalidate previous conclusions, the attack may not proceed.
- **(6) Assessment** involves evaluating the effects of the targeting process on the environment. Violations or irregularities should be monitored, measures should be taken to prevent repetition, and responsibility may be assigned (Geneva Convention I, 1949). This step requires a degree of auditability and traceability to enable these measures to be taken (Kwik & Van Engers, 2021).

As the legal tests extracted from IHL are normative in nature, we also took preliminary steps to represent these rules in a semi-formal way. We extensively explain the legal tests and motivate our exact reasons for each formulation in Kwik et al. (2022). These include inter alia *Article 57(3) Precautions* (if several targets provide comparable MA, the lower IH should be selected), *Proportionality* (decisions that

result in MA vis-à-vis an excessively high IH are illegal, with *p* representing the ratio delimiting excessive from non-excessive decisions) and *Minimisation* (from all decisions, the highest feasible MA-to-IH ratio should be selected). All legal tests are further refined below in computational terms in "The basics of the model" section, with Article 57(3) Precautions, Proportionality and Minimisation mentioned in this paragraph corresponding to Formulas 1, 2 and 3 respectively, where they will be explained in more detail.

The framework of the autonomous targeting system

In this section, we will present the structure of our autonomous targeting system that incorporates these legal tests into its functionality. Note that we will not discuss any particular targeting scenario, but rather introduce a general (function-agnostic) IHL-compliant framework. Although we will not discuss the technical details of the machinery which can be used to implement the targeting system, we propose a structure and possible techniques that allow for the creation of a targeting system which can respect and implement the legal requirements identified in previous sections. We realize that some functionalities can still be quite difficult to implement in real life systems (e.g. identifying direct participation in hostilities) Szpak (2020). However, we can expect that such modules, at least for some tasks (e.g. distinguishing military from 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 the observation that, although transparency and explainability requirements are crucial for many legal tests (Kwik & Van Engers, 2021), the requirements for the cognitive elements of the decision-making process are less restrictive. As we expand upon in "The basics of the model" section, determining to what extent a decision promotes a particular value (e.g. military benefit, civilian well-being) is sometimes straightforward, but other times highly nondeterministic. For instance, all military commanders would agree that causing 5 civilian deaths is better than causing 10: a simple formula is sufficient to capture this dynamic. What is less clear is to what extent (if at all) the death of a farmer is less grave than that of a health care worker, or to what extent destroying a museum with all its art is worse than destroying a hotel where civilians are present (Dinstein, 2016; Schmitt & Schauss, 2019). If the enemy's high command happens to hold a meeting in either of these locations, equally reasonable commanders might also arrive at differing conclusions as to whether to proceed with the attack on the basis of the proportionality rule (Wright, 2003).

IHL's solution to accommodate this inherently qualitative (Winter, 2020) exercise is to consider "the value that a reasonable attacker in the same or similar circumstances would accord" to the situation in question (Schmitt & Schauss, 2019). As such, it suffices for a commander who has chosen for the decapitation strike on the hotel to adduce the MA (= neutralising enemy high command) and IH (through collateral damage estimation) projected prior to the attack, and argue that in light of these circumstances they found the attack to be proportionate. As long as this judgment in balancing is not unreasonable, the attack would be lawful (Bartels, 2013). Like other humans, commanders would likely not be capable of explaining in granular detail the weights they attached in terms of MA and IH to each additional adjutant killed or waiter injured. An extended framework describing the factors that are to be considered when taking such decisions could be helpful to inductively discover the weighing factors from many different cases. We recognize that preferences are both, individual, circumstantial, partial orderings, impacted by cultural and religious factors etc. This is no different in military decision-making (International Criminal Tribunal for the Former Yugoslavia, 2001).

This complexity does not imply that functions of the cognitive part couldn't be created, particularly when the relationships are expressed in a straightforward way, using an explainable formula similar to the explanation commanders would give when they would prefer 5 civilian deaths above 10. The problem of the scope of transparency is a controversial one and we believe that it would require deeper discussion and clarification from legal, ethical and technical point of view. We hope to address these issues in more detail in future work.

In light of the above, and taking into account the strengths and weaknesses of the two major categories of AI-technologies (data-driven and knowledge-driven AI), we argue that the creation of an autonomous targeting system will most likely require a hybrid system, i.e. one that may contain for example ML-based classification of situations, reinforcement learning based planning modules, knowledge-based reasoning modules, etc. It cannot rely solely on rule-based mechanisms, nor on data-driven AI alone. Since the sixth stage of the targeting process (Assessment) requires evaluation and justification of why a particular decision was taken, we argue that the part of the system responsible for legal tests should best be constructed on the basis of a knowledge-based paradigm. This way, all legal tests can be performed in an explicit and transparent way, while other parts of the targeting process could perhaps be performed with the use of less transparent/explainable techniques.

The key element of the procedure described in the previous section is the comparison between anticipated MA and anticipated IH. Obviously, a human commander, while making his decisions, does not need to represent either

MA nor IH in a quantifiable form,[1] but the creation of an autonomous AI-driven model requires not only a computational model, hence a quantifiable representation, but also a representation which allows for their formal comparison.

How to represent such a phenomenon? In order to allow for comparison of both dimensions (MA and IH) we have to represent them in a form which can be used for computational analysis, especially which can be obtained with the use of various AI mechanisms. For this purpose, we are going to use *values* as a central concept allowing for representation of both MA and IH. There are number of definitions of values and approaches to modeling value-based reasoning which significantly differ in many important details. In our model we will be using the concept of values as introduced in Zurek (2017) and later (Zurek & Mokkas, 2021), where value is defined as an abstract (trans-situational) concept which allows for the estimation of a particular action or a state of affairs and influences one's behavior. According to most value-based approaches, values can be satisfied or promoted to a certain extent and they can be seen as a kind of abstraction of particular situations which allows for comparison of different values (see equations 1-3). In other words, the levels of satisfaction of particular values by a particular state of affairs (decisions with anticipated results) can be expressed by numbers and compared.

In our IHL case we represent the relevant factors MA and IH as values. Also here these values can be satisfied (or promoted) to a certain extent, which allows for representing the extent to which MA or IH are reached or obtained through a particular decision. This understanding is different from most popular approaches to reasoning with values, where values have a binary character (Bench-Capon, 2003) or can only be neutral, promoted, or demoted (Atkinson & Bench-Capon, 2016). Unlike in those other approaches (Bench-Capon, 2003), we do not introduce any fixed ordering between values. Instead, we compare the levels to which the results of particular decision will satisfy selected values. On the basis of the above (the IHL requirements in particular) we assume two main values: *Civilians*, v_{Civ} (the life, health, well-being, possessions, infrastructure of civilians) and *Military Advantage*, v_{MA}. Note that the value *Civ* is inversely proportional to the level of harm inflicted on civilians (IH).

One can notice that values like "military advantage" or "civilians" are very general and, in fact, they can be seen as the aggregation of a set of more specific values. Such a view is coherent with the Schwartz Value Theory (Schwartz, 1994). In (Zurek et al., 2022) a set of values influencing military advantage and civilians was introduced. In addition, the authors of Zurek et al. (2022) introduced a formal mechanism allowing for the calculation of the level of satisfaction of more general values (like v_{MA}) on the basis of the levels of satisfaction of more specific values (like $v_{groundGained}$, $v_{disruptiveEnemyActiv}$, etc.).

On the basis of the above, by V we denote the set of all values including, but not limited to the list from (Zurek et al., 2022). As we noted before, values can by satisfied to a particular level by an action or state of affairs. Let by $v_x(s_y)$ we denote a level of satisfaction of value v_x by a state of affairs (or action) s_y. We will not impose any particular form of representation of the level of satisfaction of a value at this point, but for the sake of this discussion, let us assume that this can be represented by a real number. Such a definition of values allows us to use them as a central concept in our model where they can play an important function as an intermediate concept representing an abstraction of a targeting situation.

On the basis of the above, we present a discussion of how from a technical viewpoint, by using the use of concept of value, the requirements of particular stages of targeting process can be fulfilled. In this paper we will present the overall structure of the system but we will not enter into the technical details of particular functions used in the model, unless this is necessary to make the model understandable or when it constitutes a key element of the discussion.

Many authors Zurek et al. (2022), Schmitt and Schauss (2019) point out that the legal tests from IHL should be performed before a final decision is made. This means that actors involved should anticipate the results of a decision to a certain extent. Such a prediction is, by definition, uncertain. Since commentators agree that uncertainty should be taken into consideration during the decision-making process (but not in every detail, see Schmitt and Schauss (2019), then for the sake of our study, we propose to use expected levels of satisfaction of values (instead of absolute levels) in the reasoning process.

The basics of the model

In "Targeting and legal tests" section, the six stages of targeting process were discussed. Below we elaborate how to represent the legal tests encountered during each of those stages:

- **Goal analysis**. In this stage the commander determines the operational objectives and desired end states as derived from strategic and operational guidance. Since such an analysis is performed from a broader perspective, taking into account the general goals of the military operation, we argue that such a goal, for the autonomous

[1] Although in more deliberate targeting settings, advanced collateral damage estimation technology has been used to provide a prognostic of IH to a high degree of accuracy (36/PAX, 2016).

device, can be represented as a set of thresholds of a group of values which constitutes the more general value *military advantage* (v_{MA}).

By representing the goals of the agent as value-thresholds, we allow the agent to autonomously state and,

when new relevant information is received, change its plan.

Since we assumed that there could be a set of values influencing *military advantage* ($v_{ground}, v_{discruptingEnemyActiv}, ...$), then the goal of an autonomous military device can be expressed as the required minimal acceptable levels of satisfaction of these (and possibly others, e.g. timing of the operation) values. Let $G = \{g_{MA}, g_{Civ}, g_{ground}, ...\}$ be a set of thresholds of values from set V representing the aim of a commander.

- **Target analysis**. This step involves the identification and specification of potential targets and whether they are legitimate military objectives.

  In order to fulfill this stage this set of preparatory tasks should be performed:

  - To generate, on the basis of signal intelligence[2] and the general state of operations, the set of potential decisions that can be made in the given circumstances. Let $S = \{s_x, s_y, ...\}$ denote the set of input vectors containing signal intelligence, general state of operations of the analysed situation, etc., and let $D = \{d_x, d_y, ...\}$ denote a set of available decisions. Suppose function $\delta : S \rightarrow D_x$ which for every $s_x \in S$ assigns a set of available decisions $D_x \subseteq D$. As previously noted, we do not introduce any particular mechanism for generating the set of available decisions (function δ). For the sake of this study we assume that creation of such a mechanism is feasible.
  - To distinguish whether the subject of the attack is a legitimate military target. Let $\theta : D \times PAR \rightarrow DD$ be a function filtering decisions which result in targeting objects or persons which may not be attacked (e.g. civilian infrastructure, persons *hors de combat*). Let $DD \subseteq D$ be a set of decisions for which unlawful targets are not the direct object of attack. For the sake of this paper we assume that function θ is granted, although in practice this requires sophisticated object/situation recognition mechanisms to realise.[3]
  - To predict the result of every decision from the set of available decisions. Note that for the tests described in "Targeting and legal tests" section, the levels of MA and IH relate to the *anticipated* results of decisions, which means that they are by nature uncertain. On the basis of that, while evaluating MA and IH, we have to take into consideration their uncertainty. If:

    * By $R = \{r_x, r_y, ...\}$ we denote a set of all possible results. In order to preserve the generality of our model, we will not discuss here how to represent the results of a decision, especially whether they should be represented by propositions, sets of parameters, etc., but this format should include information about potential injuries to civilians, civilian objects, etc.
    * By R_{s_y,d_x} we denote a set of results of decision d_x made in state s_y, which have a reasonable chance to occur.
    * By $p(r_x)$ we denote the probability of r_x, by $p(r_x \mid s_y, d_z)$ we denote a conditional probability of r_x given s_y and d_z (the probability of r_x given decision d_z in circumstances s_y)

    then by $\rho : S \times DD \rightarrow PR$, where PR is a conditional probability, we denote a partial function which returns the conditional probabilities of the results of a decision d_y made in a circumstances s_x: $\rho(s_x, d_y) = PR_{s_y,d_x} = \{p(r_t \mid s_x, d_y), p(r_k \mid s_x, d_y), ...\}$. By PR we denote a set of PRs of all available decisions.
  - To evaluate the decision results in the light of a set of relevant values.

    Suppose a set of decision results $R = \{r_x, r_y, ...\}$ and a set of functions Φ. A function $\Phi_V \in \Phi$ s.t. $\Phi_V R \rightarrow \mathbb{R}$, returns the level of satisfaction of a particular value $v_x \in V$ by result $r_y \in R$. By $v_x(r_y)$ we denote the level of satisfaction of 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. Note that although function θ excluded decisions which involve direct attacks on unlawful targets, remaining decisions may still cause *incidental* harm to such persons or objects (i.e., IH).

    Since functions from set Φ has a crucial character for our model, we briefly present how they can be obtained. There are two possible ways: (1) a particular function Φ_v can be represented in an analytical form where the level of satisfaction of value can be obtained by a formula which, on the basis of the parameters of the weapon, the number of soldiers, civilians, military and civilian objects, etc., calculates the level of satisfaction of a given value (such a mechanism is used in the current systems); (2) a

[2] "Intelligence" here is meant to refer to all necessary forms of intelligence, surveillance and reconnaissance (ISR) necessary to make reasoned targeting decisions (Curtis, 2019).

[3] Note that, as mentioned above, for some classifications, e.g. distinguishing a military aircraft from a civilian one, the task will be simpler than for others, e.g. identifying direct participation in hostilities.

particular function Φ_v can be obtained on the basis of a supervised machine learning algorithm: Suppose a set of results from set R (possible results of actions) which will be evaluated and labelled by human annotators (by assigning a number representing the level of satisfaction of a given value). This data can be used as the basis for training a regression mechanism which can predict a level of satisfaction of a given value on the basis of a particular result (Atkinson et al. (2006) introduces the mechanism of on-line democracy deliberation; a similar mechanism can be used to create the training set).

– To create a list of possible results of a given decision in a given state. For an agent in a state s_y, for every decision d_x which is available at state s_y, on the basis of function ρ the list of all results for which $p(r_t \mid s_y, d_x) > t$ should be created. t is the threshold representing the reasonable chance of a given result to occur (In case performance would be a blocking factor, we could exclude results with extremely low probabilities from the analysis).
– To calculate an expected level of satisfaction of a particular value. Let $ev_{z_{s_y,d_x}}$ denote an expected level of satisfaction of value v_z by a results of decision d_x in the state of affairs s_y:
$ev_{z_{s_y,d_x}} = \Sigma_{r_t \in R_{s_y,d_x}} p(r_t \mid s_y, d_x) v_z(r_t)$
By EV we denote set levels of satisfaction of all values by all available decisions.

The above functions allow us to distinguish the set of available decisions and derive the levels of satisfaction for all relevant values. This, in consequence, allows us to perform the ***Article 57(3)*** legal test. To recall, this provision provides that if more than one target is viable and they produce comparable MA, the target should be selected with the lowest IH. In other words, this rule requires comparison of the MA and IH of different decisions. The key point is in the rejection of the decisions (targets) which have comparable MA but can cause higher IH (in our work, the lower level of satisfaction of well being of civilians).

We will represent it by the predicate $DT(d_x)$, where d_x is the decision which satisfies the test:

$$\exists_{d_x \in D} \neg \exists_{d_y \in D} ((ev_{MA}(d_x) = ev_{MA}(d_y)) \wedge (ev_{Civ}(d_x) < ev_{Civ}(d_y))) \Rightarrow DT(d_x) \quad (1)$$

As such, the result of this test should be a set (denoted by DT) of decisions which satisfy it[4]: $DT = \{d_x \mid DT(d_x)\}$

- **Capability analysis**. During this stage an assessment of the means and methods of warfare available to the commander is performed along with two crucial legal tests: ***proportionality*** and ***minimisation of IH***.

In order to perform this step, the system requires data obtained in the previous steps (like the evaluation of MA and Civ) as well as some additional data, like the different options concerning munition, delivery, etc.

This stage includes following tasks:

– Performing the proportionality test and minimisation test:

* Proportionality test. By predicate $DP(d_x)$ we denote that decision d_x passes the proportionality test:

$$ev_{MA}(d_x) \leq p * ev_{Civ}(d_x) \Rightarrow DP(d_x) \quad (2)$$

Where p is the proportionality coefficient.
By DP we denote a set of decisions fulfilling the proportionality test: $DP = \{d_x \mid DP(d_x)\}$
* Minimisation of incidental harm. By $DMH(d_x)$ we denote that decision d_x passes the minimisation test:

$$\exists_{d_x \in D} \forall_{d_y \in D} (ev_{MA}(d_x) * ev_{Civ}(d_x) \geq ev_{MA}(d_y) * ev_{Civ}(d_y)) \Rightarrow DMH(d_x) \quad (3)$$

By DMH we denote a set of decisions fulfilling this test: $DMH = \{d_x \mid DMH(d_x)\}$

– Excluding decisions which do not fulfill commander requirements (goals and objectives) represented by thresholds of the levels of satisfaction of chosen values. This eliminates, for instance, decisions where the MA gained would be too negligible to justify an engagement, even if proportional and producing minimal IH)[5]:

In this task, the system returns the set of decisions which fulfills the assumed goals. By $DG(d_x)$ we denote that decision d_x fulfills commander requirements:

$$\forall_{v_y \in V} (ev_y(d_x) \geq g_y) \Rightarrow DG(d_x) \quad (4)$$

By DG we denote a set of decisions fulfilling the commander's requirements. $DG = \{d_x \mid DG(d_x)\}$

– Excluding decisions which do not fulfill requirements from treaties and State obligations (e.g. weapon trea-

[4] Note that v_{civ} is inversely proportional to incidental harm and we use expected levels of satisfaction of values instead of absolute ones.

[5] Note that since we also take into consideration the probability of success, we analyse the *expected* levels of satisfaction of values.

ties). This would also include restrictive (instead of prohibitive) provisions, e.g. which provide that certain weapons/ammunition may only be used under specific circumstances (Sandoz et al., 1987; Thurnher, 2014). Since in this paper we are focusing on targeting principles in IHL, we will not discuss the details of modeling and operationalizing these treaty provisions. In order to keep our model complete, however, we introduce a dedicated module responsible for filtering decisions which pass these treaty requirements. Such a task may require some additional data concerning the parameters of the decision, e.g. the type of ammunition used. In particular, it should be possible to obtain from d information concerning the type of weapon used in the attack or the types of harm the weapon may cause. Since we will not discuss these details here, for the sake of simplicity, we assume a set of parameters of decisions $PAR = \{z, t, ...\}$ (those parameters contain, for example, details of weapons used in attack). Function $\omega : D \rightarrow PAR$ returns a set of parameters of a particular decision. By $FPAR$ we denote a set of forbidden parameters of values. Predicate $DTR(d_x)$ represents that decision d_x fulfills the treaties:

$$\neg\exists_{a \in PAR}(a \in FPAR) \Rightarrow DTR(d_x) \tag{5}$$

By DTR we denote a set of decisions fulfilling the treaties: $DTR = \{d_x \mid DTR(d_x)\}$

- **Capability assignment**. This step encompasses the definitive matching of the chosen capability mix to the targets. During this stage the system should choose the decision option to be executed. This process requires analysis of not only the features of the available decision options and their predicted results, but also the probability of those results. The result if this analysis is the ordered list of decisions which:
  - Fulfill the legal requirements by passing all tests introduced in previous steps. By $DAV(d_x)$ we denote decision d_x fulfills the requirements:

$$DT(d_x) \wedge DP(d_x) \wedge DMH(d_x) \wedge DG(d_x) \wedge DTR(d_x) \Rightarrow DAV(d_x) \tag{6}$$

  By DAV we denote a set of decisions fulfilling necessary tests: $DAV = \{d_x \mid DAV(d_x)\}$
  - If no decision remains, then this means that there is no possibility to make an attack which satisfies the set military goals and which is also coherent with IHL. If there is one decision satisfying the requirements only, this decision becomes the final one. If multiple decisions satisfy the requirements, they are ordered on the basis of expected military advantage.

Let $Decisions = (D, \geq)$ be an total ordered set representing ranking of decisions. The basis of this ordering is military advantage, assuming that commanders would prefer decisions which provide the greatest military utility from all lawful alternatives:

$$\forall_{d_x, d_y \in DAV}(v_{MA}(d_x) \geq v_{MA}(d_y) \rightarrow (d_x \geq d_y)) \tag{7}$$

The structure of the system

Firstly, we assume that every function and every test presented in the previous section is performed in a specific module:

- *Extraction of available decisions* is responsible for performing function δ (S is an input, D is an output of the module);
- *Distinction module* which is responsible for performing function θ (D and PAR are inputs, while DD is an output);
- *Result prediction module* is responsible for performing function ρ (S and DD are inputs to the module, while PR and R are outputs);
- *Evaluation module* is responsible for performing function ϕ (S and R are inputs to the module, VR is an output);
- *Parameters' extraction module* is responsible for performing function ω (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 input, EV is an output of the module);
- *Treaties fulfillment module* is responsible for performing function filtering decisions which do not fulfill treaty obligations (PAR and $FPAR$ are input, DTR is an output of the module). The functioning of the module is represented by eq. 8;
- *Goals fulfillment module* is responsible for performing the function of filtering decisions which do not fulfill the commander's requirements (G and EV is input, DG is an output of the module). The functioning of the module is represented by eq. 7;
- *Harm minimisation filter* is responsible for the process of minimising incidental harm (EV is an input, DMH is an output). The functioning of the module is represented by eq. 6;
- *Proportionality test* is responsible for performing the proportionality test (EV is an input, DP is an output). The functioning of the module is represented by eq. 5;
- *Article 57(3) Filter* is responsible for the process of filtering decisions which for the same military advantage cause greater harm to civilians (Article 57(3), EV is an

input, *DT* is an output). The functioning of the module is represented by eq. 4;
- *Fulfillment of requirements* is responsible for joining together results of the above tests (*DT*, *DP*, *DMH*, *DG*, and *DTR* are input and *DAV* is an output).The functioning of the module is represented by eq. 9;
- *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 functioning of the module is represented by eq. 10.

The structure of the proposed model is presented in Fig.1. The model is created with 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. (the lower part of the scheme). Although the framework was inspired by a targeting cycle we do not replicate the procedure which has been designed for human commanders, but we adapt it to the technical requirements of the AI-based decision system: in the cognitive part we model the pipeline of preparation of data necessary to perform legal tests. In the reasoning part we represent all necessary legal tests which can be performed parallel, without the distinction for particular stages of the cycle.

The distinction between cognitive and reasoning part of the decision process is rooted in the, as mentioned earlier, assumption 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 fulfills all necessary legal tests, while other elements of the decision process can use knowledge-driven approaches. Such an approach is coherent with the general approach of hybrid systems in which data-driven parts are used as a mechanism to extract input data for the knowledge-based system, and in general allows for filling the so-called semantic gap between data and knowledge (Meyer-Vitali et al., 2019).

Another important issue which has to be discussed is the problem of dynamics of battlefield. However, although our model presents a one-moment-in-time situation, it can be understood as a constant working system which, in the case of changing circumstances (e.g. on the basis of signal intelligence), can reconsider the decision and make a new one.

Discussion and Conclusions

Although the issue of embedding legal and moral rules into autonomous devices is a popular topic of discussion, the problem of modeling a military decision-making system that explicitly incorporates IHL rules is a relatively novel one. There are a number of works (Venkatasubramanian, 2019; Schuller, 2019; Thorne, 2020) discussing this problem, but most of them are written from a legal point of view without presenting a computational solution to the problem allowing for implementation in (semi-)autonomous targeting systems. So far we have not found many papers presenting a structure of such a system with a discussion of the data necessary to perform the required legal analyses. Although (Arkin et al., 2012) presents the structure and discussion of a purely knowledge-based moral military autonomous robot, the authors focus on the moral rather than legal aspects of the decision making process. Moreover, they do not introduce any mechanism of connecting their system with data-driven approaches.

Embedding moral and legal principles into non-military autonomous devices is challenging and undoubtedly an equally important and heavily debated topic. A number of different approaches to embed legal reasoning into autonomous devices have been presented. They can be distinguished into two main research directions (Prakken, 2017): Knowledge-driven (knowledge-based) and Data-driven (ML-based). A good example of the latter approach is the data-driven approach to self-driving cars that was discussed in Webb et al. (2020). In that paper, the authors discuss Google's autonomous vehicle in the light of road safety requirements. The mechanism of following traffic rules in such a system is created on the basis of training on labelled data. Although such an approach could be very efficient, legal requirements in IHL concerning military decisions impose a number of specific tests which should be performed explicitly. Since performing such tests by a purely ML-driven device could be very difficult, we argue that the part of the system responsible for performing those tests should be knowledge-driven.

The opposite view on the problem of making decisions in a legally regulated environment is presented in the models created in the knowledge-based paradigm. Such an approach was discussed in a number of papers, for example (Prakken, 2017; Shadrin et al., 2017) (self-driving cars), or papers devoted to practical reasoning and decision making systems (Shams et al., 2016; Bench-Capon & Modgil, 2019) (and many others), BDI agents (Dignum, 1999; Meneguzzi & Luck, 2009), etc. All those approaches use logic- and argumentation-based decision making mechanisms with sophisticated inference engines. Since IHL and the targeting process imposes an ex-ante evaluation following a specific structure on the decision making process, we could take a rather straightforward model with a number of legal tests represented by logic-based formulae. Because of the ex-ante nature of the decision-making process, we did not have to consider complexity increasing reasoning methods such as reasoning with various pro and con arguments (as in the above-mentioned approaches). Obviously this advances

operational requirements, as our approach enables simple and fast testing of the available decisions.[6] This is especially important, because we assume that our model should work continuously and have a possibility to reconsider previous decisions.

Conclusions This paper introduces a framework for creating an AI-based hybrid targeting system for military autonomous agents which follows IHL rules. The main goal of our work was to introduce a discussion on the possibility of creating IHL compliant military autonomous devices. In order to enable such a discussion we have presented a construction framework of a military targeting system with the ability to follow the specific requirements provided by IHL. We present stages of the targeting process, point out which legal and moral requirements are imposed by IHL, and introduce the mechanism which allows the creation of a system fulfilling those requirements. On the basis of the framework, we can distinguish the key components and requirements necessary for the testing of compliance with IHL. Decomposition of the decision analysis process allows for a better understanding of the difficulties and challenges connected with the creation of military autonomous devices.

In our model we distinguish five legal tests, three of which are particularly important: Article 57(3) test, the harm minimisation test, and proportionality test. Those tests are important because they require not only legal analysis, but also a kind of moral evaluation of available decisions (balancing between military advantage and collateral damage). Formal models which we introduce allow for creating the reasoning mechanism (see Zurek et al. (2022) for experimental verification of some of the above tests), but the key difficulty of the implementation of these tests lies not in the reasoning or balancing process, but in the evaluation of the available decisions in the light of MA and IH (functions Φ). This constitutes the key technical and moral challenge, and it can be seen as the most controversial element of our approach, because the system should evaluate expected harmfulness of the consequences of a decision.

How can this issue be addressed? This problem can be analysed from two perspectives: (1) the evaluation will maintain a "human in the loop", i.e. a commander who will be responsible for IH and MA evaluation, or (2) the evaluation will be performed automatically by a system. The second approach is much more controversial, but it is worth noticing that even in this approach, the evaluation will also be made under the "indirect" influence of human judgement: it can be done either by explicit introduction of Φ functions (for knowledge-driven systems) or by the utilization of training data labeled by human annotators (for data-driven systems). This observation is connected to the broader question of the shape of human control: since AI-driven decision making mechanisms are made on the basis of human knowledge (knowledge-based) or data sets annotated by humans (machine learning-based), is there 'human control' over such devices? Although indirectly, humans can potentially still be viewed as having influence over the decisions made by such devices. We do not provide an answer to this question but see it as an important topic for future research.

Our 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 implementations or in what type of weapon system this framework could be incorporated, as these details would depend on the military organisation's specific needs. In a future work we plan to focus on details of selected modules of our framework. In particular, we would like to perform an experimental analysis of the possibility of obtaining of function Φ and the analysis of the mechanism of fulfilling treaty requirements.

Acknowledgements 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.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Defense Science Board: Memorandum. In: Defense science board (ed.) The role of autonomy in DoD systems. Department of Defense. (2012)

Ministère des Armées. (2019). *L'intelligence artificielle au service de la défense*. Ministère des Armées, Technical report.

Thurnher, J. S. (2018). Feasible precautions in attack and autonomous weapons. In W. H. von Heinegg, R. Frau, & T. Singer (Eds.), *Dehumanization of warfare: Legal implications of new weapon technologies* (pp. 99–117). Springer.

Abaimov, S., & Martellini, M. (2020). Artificial intelligence in autonomous weapon systems. In M. Martellini & T. Ralf (Eds.), *21st Century prometheus managing CBRN safety and security affected by cutting-edge technologies* (pp. 141–177). Springer.

Chengeta, T. (2016). Accountability gap: Autonomous Weapon systems and modes of responsibility in international law. *Denver Journal of International Law & Policy, 45*, 1–50.

[6] Note that most of the argumentation-based decision systems are computationally complex. For exemplary analysis see (Nofal et al., 2014).

Bo, M. (2021). Autonomous weapons and the responsibility gap in light of the mens rea of the war crime of attacking civilians in the ICC statute. *Journal of International Criminal Justice, 19*(2), 275–299. https://doi.org/10.1093/jicj/mqab005

Nasu, H., & McLaughlin, R. (2014). Conclusion: Challenges of new technologies for the law of armed conflict. In H. Nasu & R. McLaughlin (Eds.), *New technologies and the law of armed conflict* (pp. 247–254). T.M.C. Asser Press.

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

International Committee of the Red Cross: Ethics and autonomous weapon systems: An ethical basis for human control?, CCW/GGE.1/2018/WP. Technical report, Group of Governmental Experts of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects (2018)

Sartor, G., & Omicini, A. (2016). The autonomy of technological systems and responsibilities for their use. In N. Bhuta, S. Beck, R. Geiß, H.-Y. Liu, & C. Kreß (Eds.), *Autonomous Weapons Systems* (pp. 39–74). Cambridge University Press.

Lewis, J. (2015). The case for regulating fully autonomous weapons. *Yale Law Journal, 124*, 1309–1325.

Cummings, M. L. (2018). Artificial intelligence and the future of warfare. In M. L. Cummings, H. M. Roff, K. Cukier, J. Parakilas, & H. Bryce (Eds.), *Artificial intelligence and international affairs: Disruption anticipated* (pp. 7–18). Chatham House.

de Sio, F. S., & van dev Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI, 56*, 5. https://doi.org/10.3389/frobt.2018.00015

Eklund, A. M. (2020). *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.

Kwik, J. (2022). A practicable operationalisation of meaningful human control. *Laws, 11*(3), 43. https://doi.org/10.3390/laws11030043

Kalmanovitz, P. (2016). Judgment, liability and the risks of riskless warfare. In N. Bhuta, S. Beck, R. Geiß, H.-Y. Liu, & C. Kreß (Eds.), *Autonomous weapons systems: Law, ethics, policy* (pp. 145–163). Cambridge University Press.

Szpak, A. (2020). 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. https://doi.org/10.1017/S1062798719000310

McDougall, C. (2019). Autonomous weapon systems and accountability: Putting the cart before the horse. *Melbourne Journal of International Law, 20*(1), 58–87.

Scharre, P.D.: Army of none: Autonomous weapons and the future of war. Ww Norton & Co (2018)

Hosang, J. F. R. B. (2021). Control through ROE in military operations: Autonomous weapons and cyber operations as reasons to change the classic ROE concept? In R. Bartels, J. C. van den Boogaard, P. A. L. Ducheine, E. Pouw, & J. Voetelink (Eds.), *Military operations and the notion of control under international law* (pp. 393–420). Springer.

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

Cherry, J., & Johnson, D. (2020). Maintaining command and control (C2) of lethal autonomous weapon systems: Legal and policy considerations. *Southwestern Journal of International Law, 27*(1), 1–27.

Boothby, W. H. (2019). *New technologies and the law of war and peace*. Cambridge University Press.

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)

Adviesraad Internationale Vraagstukken. (2015). *Commissie van Advies inzake Volkenrechtelijke Vraagstukken: Autonome Wapensystemen: De Noodzaak van Betekenisvolle Menselijke Controle*. AIV.

International Committee of the Red Cross: Background paper prepared by the International Committee of the Red Cross. In: Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons, pp. 69–85. ICRC, Versoix (2016)

Organisation, North Atlantic Treaty. (2019). *Allied Joint Doctrine for the Planning of Operations*. NATO Standardization Office (NSO): Technical report.

van den Boogaard, J. C., & Roorda, M. P. (2021). 'Autonomous' weapons and human control. In R. Bartels, J. C. van den Boogaard, P. A. L. Ducheine, E. Pouw, & J. Voetelink (Eds.), *Military operations and the notion of control under international law* (pp. 421–439). Springer.

Corn, G. S. (2014). War, law, and the oft overlooked value of process as a precautionary measure. *Pepperdine Law Review, 42*, 419–466.

Curtis E. Lemay Center: Air force doctrine publication 3-60-targeting. (2019). www.doctrine.af.mil/Doctrine-Publications/AFDP-3-60-Targeting

North Atlantic Treaty Organisation: Allied Joint Doctrine for Joint Targeting, Edition A Version 1 (April 2016) AJP-3.9 (2016)

Roorda, M. (2015). NATO's targeting process: Ensuring human control over (and lawful use of) 'Autonomous' Weapons'. In A. P. Williams & P. D. Scharre (Eds.), *Autonomous systems: Issues for defence policymakers* (pp. 152–168). NATO.

Geneva Convention I: 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 (1949)

Kwik, J., & Van Engers, T. (2021). Algorithmic fog of war: When lack of transparency violates the law of armed conflict. *Journal of Future Robot Life, 2*(1–2), 43–66. https://doi.org/10.3233/FRL-200019

Kwik, J., Zurek, T., van Engers, T. (2022). Designing International Humanitarian Law into Military Autonomous Devices. https://doi.org/10.1007/978-3-031-20845-4_1

Dinstein, Y. (2016). *The conduct of hostilities under the law of international armed conflict* (3rd ed.). Cambridge University Press.

Schmitt, M. N., & Schauss, M. (2019). Uncertainty in the law of targeting: Towards a cognitive framework. *Harvard National Security Journal, 10*, 148–194.

Wright, R. G. (2003). Combating civilians casualties: Rules and balancing in the developing law of war. *Wake Forest Law Review, 38*, 129.

Winter, E. (2020). The compatibility of the use of autonomous weapons with the principle of precaution in the law of armed conflict. *The Military Law and the Law of War Review, 58*(2), 240–273. https://doi.org/10.4337/mllwr.2020.02.18

Bartels, R. (2013). Dealing with the principle of proportionality in armed conflict in retrospect: The application of the principle in international criminal trials. *Israel Law Review, 46*(2), 271–315. https://doi.org/10.1017/S0021223713000083

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. Technical report (2001). www.icty.org/en/press/final-report-prosecutor-committee-established-review-nato-bombing-campaign-against-federal

36, A., PAX: Areas of Harm - Understanding Explosive Weapons with Wide Area Effects. Article 36/PAX (2016)

Zurek, T. (2017). Goals, values, and reasoning. *Expert Systems with Applications, 71*, 442–456. https://doi.org/10.1016/j.eswa.2016.11.008

Zurek, T., & Mokkas, M. (2021). Value-based reasoning in autonomous agents. *International Journal of Computational Intelligence Systems, 14*, 896–921. https://doi.org/10.2991/ijcis.d.210203.001

Bench-Capon, T. J. M. (2003). Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation, 13*(3), 429–448.

Atkinson, K., & Bench-Capon, T. J. M. (2016). States, goals and values: Revisiting practical reasoning. *Argument Computer, 7*, 135–154.

Schwartz, S. H. (1994). Are there universal aspects in the structure and contents of human values? *Journal of Social Issues, 50*(4), 19–45.

Zurek, T., Woodcock, T., Pacholska, M., & van Engers, T. (2022). Computational modelling of the proportionality analysis under international humanitarian law for military decision-support systems. https://ssrn.com/abstract=4008946 (waiting for review)

Atkinson, K., Bench-Capon, T., & McBurney, P. (2006). Parmenides: Facilitating deliberation in democracies. *Artificial Intelligence and Law, 14*(4), 261–275. https://doi.org/10.1007/s10506-006-9001-5

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)

Thurnher, J. S. (2014). Examining autonomous weapon systems from a law of armed conflict perspective. In H. Nasu & R. McLaughlin (Eds.), *New technologies and the law of armed conflict* (pp. 213–228). The Hague: T.M.C. Asser Press.

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. Technical report, TNO (2019). TNO 2019 R11941

Venkatasubramanian, S.: Structural Disconnects between Algorithmic Decision-making and the Law (2019). blogs.icrc.org/law-and-policy/2019/04/25/structural-disconnects-algorithmic-decision-making-law Accessed 26 May 2021

Schuller, A. L. (2019). Artificial intelligence effecting human decisions to kill: the challenge of linking numerically quantifiable goals to IHL compliance. *Journal of Law and Policy for the Information Society, 15*, 105–122.

Thorne, J.G.: Warriors and War Algorithms: Leveraging Artificial Intelligence to Enable Ethical Targeting. Technical report, Naval War College (2020). https://apps.dtic.mil/sti/citations/AD1104171

Arkin, R. C., Ulam, P., & Wagner, A. R. (2012). Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception. *Proceedings of the IEEE, 100*(3), 571–589. https://doi.org/10.1109/JPROC.2011.2173265

Prakken, H. (2017). On the problem of making autonomous vehicles conform to traffic law. *Artificial Intelligence and Law, 25*, 341–363. https://doi.org/10.1007/s10506-017-9210-0

Webb, N., Smith, D., Ludwick, C., Victor, T., Hommes, Q., Favaro, F., Ivanov, G., Daniel, T.: Waymo's safety methodologies and safety readiness determinations (2020)

Shadrin, S., Varlamov, O., Ivanov, A.: Experimental autonomous road vehicle with logical artificial intelligence. Journal of Advanced Transportation **2017** (2017). https://doi.org/10.1155/2017/2492765

Shams, Z., De Vos, M., Oren, N., Padget, J.: Normative practical reasoning via argumentation and dialogue. In: Kambhampati, S. (ed.) Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI), 2016, pp. 1244–1250. AAAI Press, ??? (2016). https://doi.org/10.5555/3060621.3060794

Bench-Capon, T., Modgil, S.: Norms and extended argumentation frameworks. In: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, pp. 174–178. ACM(2019). https://doi.org/10.1145/3322640.3326696

Dignum, F.: Autonomous agents with norms. Artificial Intelligence and Law **7** (1999). https://doi.org/10.1023/A:1008315530323

Meneguzzi, F., Luck, M.: Norm-based behaviour modification in BDI agents. In: AAMAS (2009)

Nofal, S., Atkinson, K., & Dunne, P. E. (2014). Algorithms for decision problems in argument systems under preferred semantics. *Artificial Intelligence, 207*, 23–51. https://doi.org/10.1016/j.artint.2013.11.001

Zurek, T., Mohajeriparizi, M., Kwik, J., van Engers, T.M.: Can a military autonomous device follow international humanitarian law? In: Francesconi, E., Borges, G., Sorge, C. (eds.) Legal Knowledge and Information Systems-JURIX 2022: The Thirty-fifth Annual Conference, Saarbrücken, 14–16 December 2022. Frontiers in Artificial Intelligence and Applications, Vol. 362, pp. 273–278. IOS Press (2022). https://doi.org/10.3233/FAIA220479.