Can a Military Autonomous Device Follow International Humanitarian Law?

Zurek, T.; Mohajeriparizi, M.; Kwik, J.; van Engers, T.

DOI
10.3233/FAIA220479

Publication date
2022

Document Version
Final published version

Published in
Legal Knowledge and Information Systems

Legal Knowledge and Information Systems
E. Francesconi et al. (Eds.)

doi:10.3233/FAIA220479

Can a Military Autonomous Device Follow International Humanitarian Law?

Tomasz ZUREK [a,1], Mostafa MOHAJERIPARIZI [b], Jonathan KWIK [c], Tom VAN ENGERS [b]

[a] *T.M.C. Asser Institute, The Hague*
[b] *Complex Cyber Infrastructure, Informatics Institute, University of Amsterdam*
[c] *Faculty of Law, University of Amsterdam*
ORCiD ID: Tomasz Zurek https://orcid.org/0000-0002-9129-3157, Jonathan Kwik https://orcid.org/0000-0003-0367-5655, Tom van Engers https://orcid.org/0000-0003-3699-8303

Abstract. The paper presents a formal model and an experimental verification of the system controlling the International Humanitarian Law compliance for the autonomous military device.

Keywords. military autonomous device, International Humanitarian Law, reasoning model, experimental analysis

1. Introduction

Military autonomous devices remain an object of a constant debate, with the main controversies related to moral and legal issues. In particular, it is frequently argued that incorporating many principles of international humanitarian law (IHL), such as distinction, proportionality, and precautions, into an AI is impossible [1,2]. In opposition to this, other commentators [3,4] have noted that the possibility of IHL-compliant military AI should not be immediately discarded, particularly in light of the advantages a well-functioning AI can provide in the form of better performance and increased respect for the law [5,6]. In this paper we examine the possibility of implementing an IHL-compliance controlling mechanism and perform its experimental verification. On the basis of the experiment we discuss what kinds of data are required to perform necessary legal tests and point out the main difficulties of its implementation. We develop the mechanism described in [7] by introducing a fully-fledged formal representation of IHL rules and their implementation with the use of ASC2 and eFLINT languages.

[1] Corresponding Author: Tomasz Zurek, t.zurek@asser.nl. 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.

2. International Humanitarian Law rules

International humanitarian law (IHL) is the body of rules applicable to all military operations, including weapons release [8]. Attack decisions and weapons systems that do not comply with IHL principles are unlawful and may even entail the decision-maker's criminal liability for any harm that results [9]. These principles include guaranteeing that the weapon is sufficiently accurate so as to not be indiscriminate, that attacks are proportionate, and all necessary precautions are taken to spare the civilian population.

For our purposes, IHL principles related to targeting and weaponeering are particularly relevant, which are implemented through a series of legal tests during the targeting process [10,11,12]. The authors of [7] structured and streamlined these legal tests for implementation by a hypothetical military autonomous device. In the current paper, we will limit our discussion to the implementation of tests which are commonly described [4] as the most difficult tasks for an artificial agent to perform, namely those which involve the incidental harm (IH) and military advantage (MA) variables. The tests in question are the *proportionality rule* and the *two minimisation rules* (see Section 3.3).

3. The model

The general structure of our model has been presented in [7]. The key point of the model lies in the analysis of various relations between MA and IH. In our model they are expressed by two values: v_{MA} representing Military Advantage and v_{CIV} representing civilian well-being (inversely proportional to IH). In this section we introduce the three layers of the legal analysis conducted by the military autonomous device.

3.1. First layer: Data preparation

In this layer the system prepares the data needed to perform legal tests. In this paper we assume that necessary data have already been prepared. An initial discussion of the topic of obtaining this data was introduced in [7]. We realize that some functionalities may still be very difficult to implement in real life systems (e.g. identifying direct participation in hostilities) [2]. 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. Below follows the list of the data necessary to reason about the legality of a military autonomous agent's behaviour:

- The set of propositions $D = \{d_x, d_y, \ldots\}$ representing the available decisions.
- The set of evaluations of the results of decisions in the light of two values: v_{civ} and v_{MA}. The extents to which every value is satisfied by the results of the decisions are denoted by $V = \{v_{civ}(d_x), v_{MA}(d_x), v_{civ}(d_y), v_{MA}(d_y), \ldots\}$. Every evaluation is expressed by a real number.
- The proportionality coefficient p, a real number declared in advance, represents the level of acceptable (from the point of view of IHL) relations between military advantage and incidental harm to fulfil the Proportionality test.

3.2. The second layer: Weighting of MA and IH

Both MA and IH are, in our model, expressed by numbers. Since the framework requires a logical representation of norms, we have to introduce an intermediate layer of the analysis of decisions. The role of the weighting layer is to examine the relations between the levels of satisfaction of MA and IH. This problem of balancing appears in three tests in particular: (1) Article 57(3) test, (2) Proportionality test, and (3) Minimisation of Incidental Harm test. The three tests mentioned above require four kinds of weighting between v_{MA} and v_{CIV}[2]:

(1) Test whether two different decisions satisfy Military Advantage to the same level. If by d_x and d_y we denote two different decisions then by $EQMA(d_x,d_y)$ we denote that both decisions satisfy MA to the same level: $(ev_{MA}(d_x) = ev_{MA}(d_y)) \rightarrow EQMA(d_x,d_y)$

(2) Test whether one of two decisions satisfy v_{CIV} to a greater extent than the other. By $LESSCIV(d_x,d_y)$ we denote that d_x satisfies value v_{CIV} to a lower extent than d_y: $ev_{Civ}(d_x) < ev_{Civ}(d_y) \rightarrow LESSCIV(d_x,d_y)$

(3) Test whether the level of satisfaction of the well-being of civilians (v_{CIV}) by results of a given decision, multiplied by a certain coefficient, is higher than the level of satisfaction of MA by the same decision. In other words, this tests whether military advantage is proportionate to a change in the well-being of civilians. By $PROP(d_x)$ we denote that decision d_x brings about results which satisfy the test: $ev_{MA}(d_x) \leq p * ev_{Civ}(d_x) \rightarrow PROP(d_x)$

(4) Compare relations between the extents of satisfaction of MA and IH obtained by two decisions. By $MOREREL(d_x,d_y)$ we denote that the relation of MA to IH for d_x is higher than it is for d_y: $ev_{MA}(d_x) * ev_{Civ}(d_x) \geq ev_{MA}(d_y) * ev_{Civ}(d_y) \rightarrow MOREREL(d_x,d_y)$

3.3. Third layer: legal rules

On the basis of the predicates introduced in the previous section, we introduce a set of legal rules representing tests necessary to examine whether a given decision is lawful from the point of view of IHL. We use standard logical expressions to represent the above tests:

Article 57(3) test. This provision provides that if more than one target is viable and they produce comparable MA, the target with the lowest IH should be selected. We will represent it by the predicate $DT(d_x)$, where d_x is the decision which satisfies the test: $\exists_{d_x \in D} \neg \exists_{d_y \in D}(EQMA(d_x,d_y) \wedge LESSCIV(d_x,d_y)) \Rightarrow DT(d_x)$

Proportionality test. By predicate $DP(d_x)$ we denote that decision d_x passes the proportionality test: $PROP(d_x) \Rightarrow DP(d_x)$. Where p is the proportionality coefficient. $DP = \{d_x | DP(d_x)\}$

Minimisation of incidental harm. By $DMH(d_x)$ we denote that a decision d_x passes the minimisation test: $\exists_{d_x \in D} \forall_{d_y \in D}(MOREREL(d_x,d_y)) \Rightarrow DMH(d_x)$

Rule of IHL. A given targeting decision will be coherent with IHL if all the above tests are fulfilled. Therefore, on the basis of all the defined earlier predicates we can create a rule describing whether a given decision will follow IHL. By $DAV(d_x)$ we denote decision d_x fulfills the requirements: $DT(d_x) \wedge DP(d_x) \wedge DMH(d_x) \wedge DG(d_x) \Rightarrow DAV(d_x)$

On the basis of the above formulae we can distinguish a set of legal decisions, i.e. decisions which fulfill IHL rules. The decision-making system can choose one of those decisions to fulfill a military goal. A brief description of such a mechanism is presented

[2]Note that we use expected levels of satisfaction of values instead of absolute ones.

in Section 5.1 (an experimental implementation can be found on our github[3]), but full discussion of the decision making process, including the possibility of reconsideration of previous decisions, will be reserved for another paper.

4. Scenario

To test our mechanism, we compiled a scenario involving a hypothetical drone tasked with disabling an enemy's signal intelligence network, which it can achieve by targeting one of the key network points for each district. Complexity is added by introducing variables in the form of the network points' locations, added MA from neutralising enemy personnel around the target, added IH from the amount of civilian persons and objects damaged by an attack, and two types of missiles the drone can select from when engaging. We conceived three different subscenarios with different values for each of the variables, and tested whether our mechanism allows the drone to select the correct (i.e., IHL-compliant) target and means of engagement. We find that it does. For further details on this scenario and the test results, see the project's github.

5. Implementation

This section presents the basics of the implementation of the experiment. The proof of concept is implemented in two components: (1) an intentional agent that encapsulates the objectives and procedural knowledge that is implemented utilizing ASC2 framework and (2) a normative advisor that encompasses the the normative aspects i.e., rules that are implemented with ASC2 and eFLINT norms framework. The main advantage of using intentional agents and normative advisors is the separation of the analysis of legality of the decision from making the decision itself. Such a separation is important because it preserves the required level of transparency concerning the IHL compliance: in particular, it allows for clear understanding why a given decision fulfills a particular IHL rule. Since the main goal of our work is to discuss the experiments concerning the recognition whether a given decision option fulfills IHL requirements (i.e. if it is lawful an IHL perspective), we will focus on a particular element of a normative advisor (component 2), i.e. the normative reasoner, which is responsible for performing the legal tests. The normative reasoner is implemented with the use of eFLINT framework (discussed in section 5.2). Section 5.1 presents briefly the details of component 1, leaving a discussion of the complete decision process to another paper.

5.1. Intentional agents

Intentional agents are generally approached in the computational realm via the *belief-desire-intention* (BDI) model [13]. In practice, BDI agents also include concepts of *goals* and *plans*. Goals are concrete desires, plans are abstract specifications for achieving a goal, and intentions then become commitments towards plans. Our implementation was made with the use of AgentScript/ASC2 [14] language.

[3] https://github.com/mostafamohajeri/jurix2022-ihl-devices

5.2. The eFLINT norm language

The eFLINT language is a DSL designed to support the specification of (interpretations of) norms from a variety of sources (laws, regulations, contracts, system-level policies such as access control policies, etc.) [15]. The language is based on normative relations proposed by Hohfeld [16]. The type declarations introduce types of *facts*, *acts*, *duties* and *events*, which together define a transition system in which states—knowledge bases of facts—transition according to the effects of the specified actions and events.

The script defines multiple types of facts, some atomic ones like `target` and `vma`, some composite ones like `outcome` and the rest are *derived* facts. Some examples are: The fact `evciv(target,value)` which derives the expected value of civilian well-being for a target from all the possible outcomes of that target or the fact `proportionate(target)` which is derived from the proportionality formula in Section 3.3. Note that eFLINT by design includes a transition system that on every update proactively searches for all the possible facts (or acts, or duties).

6. Discussion of results

In the experiment, the list of available decisions with their evaluations is sent to the intentional agent in a sequence. After the last decision is sent, the system inspects the norms instance embedded in the advisor to see which facts are present. The results of the IHL compliance analysis are presented on the project's github. Although our eFLINT-based normative reasoning mechanism is relatively simple, the results obtained (even for controversial cases) are correct.

The problem of balancing was widely discussed in a number of AI and Law papers and legal case-based reasoning in particular. In legal CBR, the objects of comparison are either dimensions (e.g. [17]) or values (e.g. [18]). The key difference between our model and the existing ones is in the level of abstraction: both V_{MA} and V_{CIV} have a very abstract character, especially in comparison to dimensions like *number of disclosures*. An important difference also lies in the absolute representation of the level of satisfaction of values, whereas in other models of balancing, the levels of values' promotion was represented in a relative way (in comparison to other decisions, state of affairs, etc.; e.g. [19]). Moreover, in contrast to many argumentation or legal reasoning models [20,21], values in our model are not an external element of a reasoning process allowing for solving conflicts between arguments, but they are an element of a legal rule itself. The simplicity of our model, however, shows that the critical point of the reasoning process is not located in the legal reasoning, but in the calculation of the specific relations between v_{MA} and v_{CIV}. Such an observation allows us to derive a more general conclusion: the key difficulty of targeting compliance testing lies not in the legal reasoning and balancing itself, but in the process of evaluating the available options.

In practice, obtaining v_{MA} and v_{CIV} can be seen as a classification or regression task, which can be expressed as assigning numbers (representing v_{MA} and v_{CIV}) to particular decisions (represented by their specific parameters). The key question is whether the creation of such a regression mechanism is feasible at all. Answering this question will be an important topic for future research.

References

[1] Crootof R. The Killer Robots are here: Legal and Policy Implications. Cardozo Law Review. 2015;36:1837-915.

[2] Szpak A. Legality of Use and Challenges of New Technologies in Warfare – the Use of Autonomous Weapons in Contemporary or Future Wars. European Review. 2020 feb;28(1):118-31.

[3] Anderson K, Waxman MC. Law and Ethics for Autonomous Weapon Systems: Why a Ban Won't Work and How the Laws of War Can; 2013.

[4] Boothby WH. New Technologies and the Law of War and Peace. Cambridge: Cambridge University Press; 2019.

[5] Defense Innovation Board. AI Principles : Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense Defense Innovation Board. Department of Defense; 2019.

[6] Jurecic Q. Paul C. Ney Jr., General Counsel, U.S. Department of Defense, Keynote Address at the Israel Defense Forces 3rd International Conference on the Law of Armed Conflict. Lawfare. 2019 may.

[7] Kwik J, Zurek T, van Engers T. Designing International Humanitarian Law into Military Autonomous Devices; 2022. https://ssrn.com/abstract=4109286.

[8] Fleck D, editor. The Handbook of International Humanitarian Law. 3rd ed. Oxford: Oxford University Press; 2013.

[9] Additional Protocol I. 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.

[10] Ducheine P, Gill T. From Cyber Operations to Effects: Some Targeting Issues. Militair Rechtelijk Tijdschrift. 2018;111(3):37-41.

[11] von Heinegg WH. Considerations of Necessity under Article 57(2)(a)(ii), (c), and (3) and Proportionality under Article 51(5)(b) and Article 57(2)(b) of Additional Protocol I. In: Kreß C, Lawless R, editors. Necessity and Proportionality in International Peace and Security Law. Oxford: Oxford University Press; 2020. p. 325-42.

[12] Corn GS. War, law, and the oft overlooked value of process as a precautionary measure. Pepperdine Law Review. 2014;42:419-66.

[13] Rao AS, Georgeff MP. BDI Agents: From Theory to Practice. In: Proceedings of the First International Conference On Multi-Agent Systems (ICMAS-95); 1995. p. 312-9.

[14] Mohajeri Parizi M, Sileno G, van Engers T, Klous S. Run, Agent, Run! Architecture and Benchmarking of Actor-Based Agents. New York, NY, USA: Association for Computing Machinery; 2020. p. 11–20.

[15] van Binsbergen LT, Liu LC, van Doesburg R, van Engers T. EFLINT: A Domain-Specific Language for Executable Norm Specifications. In: Proceedings of the 19th ACM SIGPLAN International Conference on Generative Programming: Concepts and Experiences. New York, NY, USA: Association for Computing Machinery; 2020. p. 124–136.

[16] Hohfeld WN. Fundamental Legal Conceptions as Applied in Judicial Reasoning. The Yale Law Journal. 1917;26(8):710-70.

[17] Bench-Capon TJM, Atkinson K. Dimensions and Values for Legal CBR. In: Wyner AZ, Casini G, editors. Legal Knowledge and Information Systems - JURIX 2017: The Thirtieth Annual Conference, Luxembourg, 13-15 December 2017. vol. 302 of Frontiers in Artificial Intelligence and Applications. IOS Press; 2017. p. 27-32.

[18] Bench-Capon T, Prakken H, Wyner A, Atkinson K. Argument Schemes for Reasoning with Legal Cases Using Values. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law. ICAIL '13. New York, NY, USA: ACM; 2013. p. 13-22.

[19] Grabmair M. Predicting Trade Secret Case Outcomes Using Argument Schemes and Learned Quantitative Value Effect Tradeoffs. ICAIL '17. New York, NY, USA: Association for Computing Machinery; 2017. p. 89–98.

[20] Bench-Capon TJM. Persuasion in Practical Argument Using Value-based Argumentation Frameworks. Journal of Logic and Computation. 2003 06;13(3):429-48.

[21] Zurek T, Araszkiewicz M. Modeling teleological interpretation. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law. ACM; 2013. p. 160-8.