Values, Proportionality, and Uncertainty in Military Autonomous Devices

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

DOI
10.1007/978-3-031-58202-8_13

Publication date
2024

Document Version
Final published version

Published in
Value Engineering in Artificial Intelligence

Values, Proportionality, and Uncertainty in Military Autonomous Devices

Tomasz Zurek[1,2(✉)], Jonathan Kwik[1,3], and Tom van Engers[2,3,4]

[1] T.M.C. Asser Institute, University of Amsterdam, Amsterdam, The Netherlands
{t.a.zurek,h.c.j.kwik}@uva.nl
[2] Complex Cyber Infrastructure, Informatics Institute, Faculty of Science, University of Amsterdam, Amsterdam, The Netherlands
t.m.vanengers@uva.nl
[3] Faculty of Law, University of Amsterdam, Amsterdam, The Netherlands
[4] TNO, The Hague, The Netherlands

Abstract. This position paper presents a discussion on the problem of implementing the rules of International Humanitarian Law in AI-driven military autonomous devices. Fulfilling the proportionality test is one of its key requirements. This test, in order to exclude attacks causing excessive collateral damage, requires comparison of two values: anticipated military advantage and expected incidental harm. In this paper we introduce a discussion of how to take into consideration the problem of uncertainty inherent to all military attacks, and how to combine it with the evaluation process.

Keywords: Model · Proportionality rule · Uncertainty · autonomous device

1 Introduction

We are witnesses of the rapid development of autonomous devices in recent years. Although the existence of robot vacuum cleaners do not seem to pose serious risks to humans, more complex devices, like autonomous cars, may prove dangerous not only for their users, but also for bystanders such as other traffic participants. Moreover, following [22], we claim that the increasing autonomy of devices requires much more than specific limitations of their "freedom" of conduct (like Asimov's famous rules of robotics) and calls for moral or ethical reasoning that should be a crucial internal element of their entire decision process. Of all autonomous devices, those intended for military purposes are among the most controversial and potentially dangerous, a concern that has given rise to a large body of ethical, policy, and legal literature on the subject [12,13]. Legally, there is consensus [15,28] that such devices must in any event conform to International Humanitarian Law (IHL) rules in force.

Tomasz Zurek received funding from the Dutch Research Council (NWO) Platform for Responsible Innovation (NWO-MVI) as part of the DILEMA Project and from TRUST RPA project at the University of Amsterdam.

N. Osman and L. Steels (Eds.): VALE 2023, LNAI 14520, pp. 219–236, 2024.
https://doi.org/10.1007/978-3-031-58202-8_13

In [18,37,38] a model of an IHL-compliant hybrid decision-making mechanism for military autonomous devices has been introduced. Since one of the key requirements imposed by IHL is the capability of explaining compliance with the law [17] and most of the machine-learning mechanisms used in autonomous devices (especially deep learning neural networks) lack sufficient explainability in terms of the legal and ethical rules applied, the authors of [18,37,38] introduced a hybrid mechanism which includes both data- and knowledge-driven approaches. The data-driven models are responsible for generating the list of available decisions, predicting their results, and (supported by a knowledge-based system) evaluation of decisions in the light of necessary values, while the knowledge-based part is responsible for performing legal tests.

Legal literature [14,25,27] makes frequent reference to the probabilistic dimension of military decision-making: a commander not only needs to take into consideration the predicted results of a given decision, but also the *probability* of this decision bringing about said result. Although the model presented in [37] assumes that legal tests require expected levels of satisfaction of relevant values (taking into consideration also probability of success), this model does not present any details of such a machinery.

In this paper we make preliminary steps toward filling this gap by discussing alternative approaches for dealing with uncertain knowledge for military autonomous devices using the mechanism from [37]. Since most data-driven approaches (including complex machine learning models) are constructed on the basis of statistical analysis, we explore approaches that may function as a kind of bridge between knowledge- and data-driven approaches. The model presented in [37] was created on the basis of the NATO Standard Targeting Procedure [19,21]. In this paper, however, we are abstracting from the technical details concerning the data quality, source of data for the statistical analysis, etc. which should be an object of specific regulations. Moreover, we assume for the purposes of this paper that the device is capable of correctly recognising objects, predicting the results of decisions made, and evaluating them in the light of values. Note that the actual feasibility of such functionalities will depend on the task and the state of the surrounding technology.[1]

It is important to note that in this paper we will not introduce a comprehensive model for processing uncertainty in proportionality analyses, but rather sketch possible development directions and open discussion for future work. Although in this paper we focus on military autonomous devices and IHL, note

[1] The difficulty in executing these functions will be very context-dependent. For example, existing technologies are already used for distinguishing civilian from military aircraft, but distinguishing combatants from civilian persons is much more challenging. As such, some commentators have proposed prohibiting the use of autonomous targeting for these 'difficult' domains (e.g. [23]). The topic of autonomous military devices is undoubtedly controversial, and a detailed discussion of some moral issues related to this topic was presented in [35,37,38]. We take no further position related to *specific* systems or tasks: in cases where the device is not capable of performing these required functions, it simply should not be used.

that such mechanisms could equally be used in other hybrid AI-based decision making mechanisms.

2 Proportionality Analysis

2.1 The Rule

The proportionality rule [2,10] is a fundamental precautionary test in IHL that all parties conducting an attack must apply. The rule provides that belligerents must "refrain from deciding to launch any attack which may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated" [5].

As evidenced by the words 'expected' and 'anticipated', the proportionality test involves strong prognostic elements and epistemic insecurity [14,20], making it an ideal proving ground to examine how military autonomous devices would deal with uncertainty and probabilities. Uncertainty is a factor in many other IHL rules [2,27], but for the current paper, we shall focus on the proportionality test as a proof-of-concept.

The element of uncertainty in the proportionality test itself mainly attaches to two variables: 'expected incidental loss' (IH) and 'anticipated military advantage' (MA). The rule requires decision-makers to compare the two, and an attack may only proceed if the former is not excessive to the latter. Obviously, while making their decisions, a human commander does not represent IH or MA in quantifiable form.[2] However, when this test is conducted by an autonomous device, this requires not only a computational model, hence a quantifiable representation, but also a representation which allows for their formal comparison.

It is also important to emphasize that the proportionality test is only a 'floor' criterion: as long as the threshold is met, all 'remaining' decisions are lawful *under the proportionality rule*, but may still breach other targeting requirements [24,26]. In other words, the proportionality test's output is not the final outcome of the overall engagement decision: the final decision is also influenced by complementary IHL rules (such as minimisation of IH) that further reduce risk to civilians [5]. In contrast, if no decision overcomes the proportionality threshold, then no action is permitted, making the proportionality test nevertheless crucial.

2.2 The Model

To represent the process of proportionality analysis we use the model introduced and discussed in [18,37,38].

In order to allow for comparison of both dimensions (MA and IH) we need to represent them in a form which can be used for computational analysis, especially

[2] 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 [4].

which can be obtained with the use of various AI mechanisms. For this purpose, we will use *values* as a central concept allowing for representation of both MA and IH. There are a number of definitions of values and approaches to modeling value-based reasoning which significantly differ in many important details. In our model we will use the concept of values as introduced in [36] and later [39], 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 Eqs. 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 is achieved or IH is caused through a particular decision. This understanding is different from most popular approaches to reasoning with values, where values have a binary character [8] or can only be neutral, promoted, or demoted [7]. Unlike in those other approaches [8], we do not introduce any fixed ordering between values. Instead, we compare the levels to which the results of particular decisions will satisfy selected values. Most of the existing approaches to practical reasoning with values [6,33] use values to evaluate changes of states, while in our approach values are used to evaluate states (consequences) itself (i.e. [34,36,40]). Such an approach is more suitable for representing the proportionality rule, because it requires evaluation of the action's consequences, rather than changes of states. Our model is created on the basis of the approach presented in [18,40] in which values are used to evaluate the decisions' consequences in order to exclude the immoral (illegal). The treatment of values in [34] is somewhat similar to our approach in that they are abstract properties associated with propositions, but this model is focused on the decision making process, not the filtering out of unacceptable decisions.

On the basis of the above (and 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). In order to obtain v_{MA} and v_{Civ} [37] introduces a set of functions:

– To generate, on the basis of signal intelligence[3] 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 $\delta : S \rightarrow 2^{D_x}$ which for every $s_x \in S$ assigns a set of available decisions

[3] "Intelligence" here is meant to refer to all necessary forms of intelligence, surveillance and reconnaissance (ISR) necessary to make reasoned targeting decisions [11].

$D_x \subseteq D$. 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 predict the result of every decision from the set of available decisions. Note that the levels of MA and IH relate to the *"anticipated"* and *"expected"* 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 C is a set of propositions representing consequences of actions, by $\Psi : S \times D \rightarrow 2^C$ we denote a function which maps consequences to decisions made in particular circumstances. For an agent in a state s_y, for every decision d_m which is available at state s_y, the set C_m represents a set of sets of all possible results of decisions. Since not every result may have the same probability of occurrence, by $\rho : S \times D \times C \rightarrow PR$, we denote a partial function which returns the conditional probabilities that the $c_t \in C$ will be result of a decision d_y made in a circumstances s_x.
- To evaluate the decision results in the light of a set of relevant values. Suppose a set of decision results $C = \{c_x, c_y, ...\}$ and a set of functions Φ. A function $\Phi_V \in \Phi$ s.t. $\Phi_V : C \rightarrow \mathbb{R}$, returns the level of satisfaction of a particular value $v_x \in V$ by result $c_y \in R$. By $v_x(c_y)$ we denote the level of satisfaction of value v_x by result c_y, by VR we denote a set of levels of satisfaction of all values by the results of all available decisions.
  Since functions from set Φ have 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 particular function Φ_v can be obtained on the basis of a supervised machine learning algorithm: Suppose a set of results from set C (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.
- To calculate an expected level of satisfaction of a particular value. Let $ev_z(d_x)$ denote an expected level of satisfaction of value v_z by a results of decision d_x in the state of affairs s_y. The calculus will be discussed in Sect. 4.2.

Interestingly, such an approach combines the probability of particular consequence of decision with its evaluation in the light of one of the values (v_{MA}, v_{Civ}). This allows us to evaluate a particular decision not only in the light of its consequences but also in the light of the chance of their occurrence. Moreover, it can be also understood as a kind of consequentialist approach in which we evaluate the anticipated consequences of actions (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 proportionality test. By predicate $DP(d_x)$ we denote that decision d_x passes the proportionality test,

$$ev_{MA}(d_x) \leq p * ev_{Civ}(d_x) \Rightarrow DP(d_x) \quad (1)$$

where p is the proportionality coefficient.[4]

One of the important problems is how to calculate the necessary probabilities and expected levels of satisfaction of the values.

3 Uncertainty in MA and IH Evaluation

The formal model presented in [37] supports the tests required by IHL, i.e. the evaluation of the level of satisfaction of two values: MA and IH by results of such a decision for every available decision option. Since the prediction of results of a decision is never 100% certain, it is necessary to use expected levels of satisfaction of values.

Suppose that we have a set of possible decisions D. Let by c we denote a single atomic result of a given decision (for example destroying an enemy tank). One decision can cause a set of atomic results, each with its own probability (for example, firing a missile in a certain direction can cause two atomic results: destroying an enemy tank with probability 0.8 and damaging a civilian building nearby with probability 0.6). On this basis, every decision can bring about a whole spectrum of different combinations of results with different probabilities.

The key difficulty lies in the estimation of the probability of the expected results of a given decision. In other words, how we can construct a result set that contains all foreseeable results of potential decisions the device (or a human in- or on-the-loop for that matter) could take with their corresponding probability.

In this paper we discuss two issues. The first is how to construct a weighing procedure that enables a device or human to select an optimal decision from a set of possible decisions that satisfy the conditions set by the proportionality rule. And second, to find a suitable calculus for calculating the probabilities that feed into the weighing procedure.

What matters for the weighing exercise envisioned by the proportionality rule is not the actual MA achieved, which can only be known ex-post, but rather its anticipated scope, i.e. ex-ante [20]. Opinions on how to account for the likelihood of the MA materializing (or simply put, the attack succeeding) vary [9], but it is generally accepted that "the 'concrete and direct advantage anticipated' is not the value of the target wholly in the abstract but rather its abstract value relative to the likelihood of in fact neutralizing or destroying the object." [1] The degree of uncertainty of IH occurring, however, does not need to be factored in:

[4] We do not introduce any mechanism for calculating the proportionality coefficient. In our view, due to very specific requirements and circumstances of every military operation, such a coefficient should be declared by a commander before a mission. The declarative character of this coefficient can be seen as an element of human control over the device.

"[O]nce [IH] is expected, it must be calculated into the proportionality analysis as such; it is not appropriate to consider the degree of certainty as to possible [IH]." An opposing view can be found in [3,9,25], where the authors argue that uncertainty is necessary to represent both values.

Although the first stance is a dominant one in legal scholarship, from a computational point of view it can be seen as controversial, as it compares the ostensibly certain IH with uncertain MA. This different stance could be the result of misunderstanding of various concepts expressing uncertainty, for which different terms, sometimes with subtle differences in meaning are used, such as likelihood, probability, chance, odds etc.

The key point is in the understanding of the decision's uncertainty. In the model this uncertainty is factored in the formula, but we have not discussed yet what the most appropriate calculus for that (un)certainty is. In the model presented in Sect. 4.2 we formalized the conditional probability of c_x given s_y and d_z (the probability of c_x given decision d_z in circumstances s_y) as $p(c_x \mid s_y, d_z)$.

However we can distinguish here at least two levels of uncertainty. First, epistemic uncertainty regarding the fidelity of input data and basic assumptions that feed into the decision procedure,[5] and second, whether a particular decision would bring about a certain consequence. Take a decision to kill a high-level enemy leader. With regard to this attack, there may be uncertainty with regard to the leader being present at a particular location at a specific time, and uncertainty with regard to whether the chosen action (e.g. attack the tavern where he was spotted) does achieve the desired effect (i.e., the leader being killed[6]). Obviously there is a dependency between the uncertainty of the input leading to the decision and the uncertainty about the success of the taken action.

Additionally, we can also distinguish another level of uncertainty, the uncertainty of evaluation. This is independent from the uncertainty of results, but depends rather on the quality of function Φ mapping results to the values to be promoted or demoted. In our initial model we did not take these uncertainty factors nor a suited calculus that would allow for combining uncertainty and use uncertainty in a meta-reasoning about the results of applying our decision-making model to some scenario.

On the basis of the above we can distinguish the following levels of uncertainty:

- Source uncertainty: Describes uncertainty in data that are factors in the decision-making model, and hence impact the (un)certainty of the decision. Although intuitively the probabilities of data may be obtained on the basis of a statistical analysis of the sources and many researchers use Bayesian statistics for the representation of such an (un)certainty (represented as $p(s_y)$),

[5] Sensor data, signals intelligence, human intelligence, etc. While some sources are more dependable that others, none are 100% certain. In addition, processing such data (e.g. interpreting image or video feeds, whether done by humans or AI-driven systems) introduce an additional probability of mistakes [29].

[6] Note that IH is not dependent on MA: even if the leader is not present or not successfully killed, the damage to the tavern and civilians inside will remain.

there are some basic assumptions underlying Bayesian statistics that may not hold generically.
- Uncertainty of prediction of the decision results: How (un)certain are the predictions concerning the results of actions: chances of destroying a particular object, destruction to civilian objects, etc. Such an (un)certainty depends on the type of weapon, its precision, range, environmental circumstances, surroundings, anticipated number of proximate civilians, etc. (Un)certainty of prediction is much more complex to be obtained statistically because of the lack of data (both quantitatively as well as qualitatively, i.e. there may be many factors determining the effect of an action) to perform such statistical analysis. Some authors propose to represent relations between decisions and results as Bayesian nets describing particular scenarios (see e.g. [32]). In [37] this probability was represented by the formula: $p(c_x \mid s_y, d_z)$ without the discussion of how to calculate these probabilities. Also here some basic assumptions underlying Bayesian statistics may not hold.
- Evaluation uncertainty: The uncertainty that a given results of a decision will be correctly evaluated in our case in the light of values v_{MA} and v_{Civ}.

All these three uncertainties should be taken into consideration while calculating expected levels of satisfaction of v_{MA} and v_{Civ}. There is, as stated before, an open question which calculus should be used to obtain a numerical representation of uncertainty. And although some of the basic assumptions underlying Bayes may not hold, Bayesian (or naive Bayesian) networks may still be accurate enough to be useful in practice to come to reasonable results as alternative approaches may also have their limitations.

4 The Model

In this section we discuss how to formalize the uncertainty issues listed in the previous section in order to extend the current decision-making model that includes the proportionality analysis required by IHL. Since source and evaluation (un)certainty could in principle be obtained from statistical data, in this paper we focus on the uncertainty of the decisions, rather than on the (un)certainty of the decision process' input data or that of the effects of the action(s) that result from these decisions.

4.1 Model Limitations

Before we introduce the model, we introduce some simplifications:

- We exclude the influence of other agents whose decisions and consequent actions may impact the actual situation, and hence the consequences of the decisions and consequent actions of the agent in scope. For now, we assume that the possible influence of other agents is implicitly represented in the uncertainty of decision consequence pairs.

– We assume that a decision brings about a set of consequences represented by propositions. We realize that these consequences may not occur simultaneously nor may not be completely independent.
– We assume that a particular decision in a particular circumstances may result in a set of different consequences with possible different uncertainty (functions ρ and Ψ). Note that in such an understanding we do not take into consideration continuous variables (for example, after the attack on the fuel magazine the enemy will lose from 0 to 100000 liters of fuel).

4.2 Probabilistic Approach

Every $c_x \in C$ represents a particular atomic result of an action, for example the number of civilians killed, the destruction of a school building, the killing of an enemy commander, and so on.

By $v_{MA}(c_x)$ we denote a level to which atomic consequence satisfies value v_{MA}. Likewise, by $v_{Civ}(c_x)$ we denote a level to which an atomic consequence satisfies value v_{Civ}. In Sect. 3 we introduced various levels of uncertainty. One of these levels is evaluation uncertainty, the uncertainty of the process of evaluation of decision consequences in the light of particular values. In our model this evaluation is represented by a function from set Φ. For the sake of simplicity we assume that this function returns 100% certain results, leaving the discussion of evaluation uncertainty to future work.

If by $S = \{s_a, s_b, ...\}$ denote the set of input vectors containing signal intelligence, general state of operations of the analysed situation, etc., then by $p(s_a)$ we denote a probability that s_a correctly represents an actual situation.

By $p(c_x|s_a, d_m)$ we denote the conditional probability of occurring of consequence c_x when the decision d_m is made (obtained with the use of function Ψ.

Our goal is to calculate the expected levels of satisfaction of given values by a given decision (e.g. $ev_{MA}(d_m)$ for value military advantage by decision d_m and $ev_{Civ}(d_m)$ for civilians by decision d_m).

Basic Approach. By C_m^a we denote a set of atomic consequences of decision d_m that occur together as one joint consequence (denoted by a) of a given decision.[7] By C_m we denote a set of sets representing all possible joint consequences of a decision d_m. If we assume that all available propositions have non-zero probability of being the consequence of d_m then set C_m will contain $2^{|C|}$ sets of propositions. Following the above, it is possible that a particular atomic consequence can be element of two different joint consequences, for example it can be a case in which $c_x \in C_m^a \wedge c_x \in C_m^b$, which represents that two different decisions

[7] Note that in reality those consequences may not occur simultaneously nor may we assume complete independence here. One can imagine that the destructing impact of a piece of material of some mass travelling with some speed also depends on the shock wave, that may travel a slower speed, consequently arriving at the object of impact.

cause the same atomic consequence (for example a number of civilian deaths), but other atomic consequences in both sets can be different.

By $p(C_m^a|s_x, d_m)$ we denote a probability of occurrence of a joint consequence C_m given s_x after decision d_m.

The level of satisfaction of value v_{Civ} by a joint consequence of decision d_m is equal: $v_{Civ}(C_m^a) = \Sigma_{c_\alpha \in C_m^a}(v_{Civ}(c_\alpha))$

The expected level of satisfaction of a value v_{Civ} by joint consequences of decision d_x in a given circumstances s_x is equal:

$$ev_{Civ}(d_m) = \Sigma_{C_m^\alpha \in C_m}(v_{Civ}(C_m^\alpha)p(C_m^\alpha|s_x, d_m)) \tag{2}$$

Analogically:

$$ev_{MA}(d_m) = \Sigma_{C_m^\alpha \in C_m}(v_{MA}(C_m^\alpha)p(C_m^\alpha|s_x, d_m)) \tag{3}$$

4.3 Discussion and Further Development

The model presented in Sect. 4.2 represents a very general approach to uncertainty of prediction of the decision results. The key question here is how to obtain a conditional probability $p(C_m^a \mid s_x, d_m)$. Note that C_m^a is the set of propositions representing consequences of the actions and the probability of a particular elements of C_m^a, i.e.: $p(c_a \mid s_x, d_m)$ and $p(c_b \mid s_x, d_m)$ s.t. $c_a, c_b \in C_m^a$, do not have to be the same, they do not have to occur simultaneously, they can also be conditionally dependent, moreover due to a potentially huge number of elements of C_m, the number of such probabilities can be also very high. All these properties make the process of obtaining them extremely difficult, especially when there is lack of sufficient data allowing for statistical analysis.

On the basis of the above we have to introduce some additional simplifications. One of the approaches will be to decompose a single C_m^a into set of propositions. How, in such a case, can joint conditional probability $p(C_m \mid s_x, d_m)$ be calculated? There are multiple ways to perform this task:

- The simplest approach is to assume that all propositions in C are conditionally independent. We can call it a naive approach (similar assumption to a naive Bayes classifier). In such a case the calculation of the joint probability of consequences will be relatively easy:
  $p(C_m^a \mid s_x, d_m) = \Pi_{c_\alpha \in C_m^a} p(c_\alpha \mid s_x, d_m) * \Pi_{c_\beta \notin C_m^a}(1 - p(c_\beta \mid s_x, d_m)$.
  The above approach, however, has a very strong disadvantage concerning the conditional independence assumption, because in a real life military situations, such dependencies may play important role. Imagine an enemy munition magazine surrounded by a group of enemy tanks. If a missile hits the magazine the probability of also destroying tanks will be much higher than if the missile hits a tank or building nearby.
- An alternative approach would be to represent C_m^a as a scenario expressed by a Bayesian network. A Bayesian network is a directed acyclic graph with probability tables for each node in the graph. Each node in the network represents a variable that can have several values e.g. true/false (more than two

values are also possible, but for the sake of simplicity we will use only a 2 value system). The probability table of a node V gives the conditional probabilities for that node taking each value given the values of its parents. The example of similar approach, but to model criminal scenarios, can be found in [32]. The basic idea is to represent scenario as a Bayesian network, where nodes represents particular events in the scenario and arrows represent causal relations (however technically they represent possible correlations only). Due to space limitations we will not present here a full introduction to Bayesian networks, but rather we will focus on a general approach to the problem. A full description of the model will be presented in a separate paper, but this is future work.

5 Example

Below we present a scenario modified from [38] on the basis of which we are going to test the extended version of our original mechanism:

> A commander from nation Alpha is given the task to disrupt enemy command and control in a city defended by nation Beta. Destroying data-centers (facilities used by Beta to collect intelligence) and neutralizing Beta's high-ranking officers will both aid Alpha in achieving this goal. Alpha's commander releases *Cleopatra* drones to attack these data-centers and Beta officers. The drones are able to identify civilians and enemy soldiers in and around potential target locations and take this information into consideration for their decision-making (the risk of misidentification or released munitions missing the target is negligible). During this operation, a *Cleopatra* identifies a data-center where Beta's general is also currently hiding. *Cleopatras* carry two types of ammunition, 'light' and 'heavy' missiles. Data-centers can be disrupted by attacking their roof-mounted antennae with light missiles, but targeting officers inside buildings usually requires the greater destructive power of heavy missiles. The heavy missile, however, is also likely to kill 150 civilians that are residing in the same building. Evidently, due to its higher power, the heavy missile is also (marginally) more likely to successfully disrupt the data-center.

This scenario simulates a complicated situation where a lower-value target (the data-center) is co-located with an extremely high-value target (Beta's general), but his presence is connected with the presence of approximately 150 civilians, giving possible decisions to examine. The question is whether the decision to fire a heavy or light missile into this building will be proportional. For the sake of simplicity we assume that we are 100% sure about the input data and the situation description ($p(s_x) = 1$). Figure 1 presents the causal dependencies between events on the basis of which the Bayesian network has been created (Tables 1 and 2).

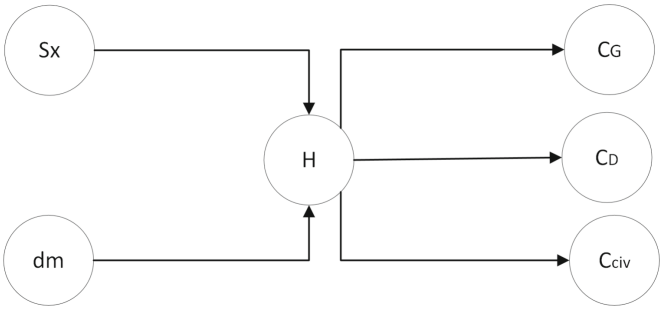

Fig. 1. Causal dependencies

Events:

- Decision d_m concerning firing the heavy missile (for the sake of analysis we assume that when the decision is made the missile will be fired);
- Decision d_n concerning firing the light missile (for the sake of analysis we assume that when the decision is made the missile will be fired);
- Hitting the target (H). Probability: $p(H \mid s_x, d_m) = 0.9$;
- Killing the general (C_G). Probabilities in Tables 1 and 2;
- Disrupting the data-center (C_D). Probabilities in Tables 1 and 2;
- Killing 150 civilians (C_{civ}, for the sake of simplicity we described it as a one event). Probabilities in Tables 1 and 2;

Table 1. Conditional probabilities of a decision d_m (heavy missile) consequences

Given	$p(C_G \mid H/\neg H)$	$p(C_D \mid H/\neg H)$	$p(C_{civ} \mid H/\neg H)$
H	0.8	0.9	0.8
$\neg H$	0.01	0.01	0.01

Table 2. Conditional probabilities of a decision d_n (light missile) consequences

Given	$p(C_G \mid H/\neg H)$	$p(C_D \mid H/\neg H)$	$p(C_{civ} \mid H/\neg H)$
H	0.1	0.9	0.2
$\neg H$	0.01	0.01	0.01

Every combination of consequences of decisions d_m and d_n is a separate set, subset of C_m. The levels of satisfaction of values v_{MA} and V_{Civ} by sets representing decision's consequences are presented in Table 3.[8]

[8] We assume that the evaluations of various consequences in the light of the values are given. We realize, however, that obtaining this data is a very difficult and complex task. See [37,38] for a discussion of the possibilities of obtaining the evaluations.

Table 3. Evaluation of decisions' consequences in the light of values and probabilities calculated on the basis of bayesian network

C_m	Consequences	v_{MA}	v_{Civ}	$p(d_m)$	$p(d_n)$
C_m^1	$C_G \wedge C_D \wedge C_{Civ}$	0.98	0	0,43525	0,01532
C_m^2	$\neg C_G \wedge C_D \wedge C_{Civ}$	0.8	0	0,10209	0,12398
C_m^3	$C_G \wedge \neg C_D \wedge C_{Civ}$	0.9	0	0,04304	0,00151
C_m^4	$C_G \wedge C_D \wedge \neg C_{Civ}$	0.98	0.8	0,10209	0,05764
C_m^5	$\neg C_G \wedge \neg C_D \wedge C_{Civ}$	0	0	0,01009	0,01226
C_m^6	$\neg C_G \wedge C_D \wedge \neg C_{Civ}$	0.8	0.8	0,02395	0,46643
C_m^7	$C_G \wedge \neg C_D \wedge \neg C_{Civ}$	0.9	0.8	0,01010	0,00570
C_m^8	$\neg C_G \wedge \neg C_D \wedge \neg C_{Civ}$	0	0.8	0,00237	0,04613

Note that killing the general and disrupting a data-center results in a promotion of a v_{MA}, while killing civilians results in demotion of v_{Civ}. For the sake of simplicity we assume that all evaluations are 100% certain.[9]

On the basis of the above and the formulas 2 and 3 we can calculate $ev_{Civ}(d_m)$, $ev_{MA}(d_m)$, $ev_{Civ}(d_n)$, and $ev_{MA}(d_n)$:

$ev_{Civ}(d_m) = 0,12253$
$ev_{MA}(d_m) = 0,632019$
$ev_{Civ}(d_n) = 0,47680$
$ev_{MA}(d_n) = 0,58133$

The above allows us to calculate whether a decision d_m will pass the proportionality test. In order to do that we need to assume a coefficient p.[10] If we assume that $p = 2$, then on the basis of formula 1 since:

for d_m: $0,632019 \not\leq 2 * 0,12253$,
for d_n: $0,58133 \leq 2 * 0,47680$,

then decision d_m (the heavy missile) does not pass the proportionality test due to the very high expected collateral damage, and the drone cannot proceed with this attack. In contrast, decision d_n (the light missile) fulfills the proportionality rule's requirements. The drone would thus use the light missile to disrupt the data-center, but leave the enemy general (and the civilians) unharmed.[11]

6 Discussion

Currently, we have two main approaches in AI, data-driven and knowledge-driven AI, that both have their strengths and weaknesses. The knowledge-driven app-

[9] In other cases, all evaluations should be multiplied by their respective probabilities.

[10] As previously noted, the proportionality coefficient p, a real number declared in advance, represents the level of acceptable (from the point of view of IHL) relationship between MA and IH.

[11] This would also be the reasonable (legally correct) decision if this scenario were presented to a human commander.

roach enables us to have clear relationships to legal sources, using some reference mechanism, that can be used for impact assessment of changes in these sources and in case interpretations of the norms in these sources change, but also in explaining decisions in terms of those sources. The downside of this approach is the problem to reason about incomplete and/or uncertain data. On the other hand, data-driven approaches have an advantage that they can be used to reason even about uncertain or missing data, but they have clear issues with explainability. A hybrid solution, combining the two would possibly overcome these issues. In the recent years there were some attempts to model probabilistic reasoning in a legal context (see e.g. [30–32]), but the work presented in literature doesn't make a clear distinction between institutional reasoning (about the institutional rules) and reasoning about social reality. In our work focusing on making decisions compliant with international law, the requirement concerning explainability of decisions [17] limits the possibilities of utilisation of purely data-driven approaches. In the research presented in this paper we tried to get to a best of breed solution. International Humanitarian Law and the proportionality rule which we discuss in this paper are particularly interesting, because they directly refer to (and require reasoning with) abstract values like military advantage or the well-being of civilians, and also acknowledge the problem of uncertainty in decision-making.

Although IHL points out that decision makers should take uncertainty into consideration,[12] it does not provide us with any specific guidance helps to deal with such uncertainty. The latter would be needed to develop practical tools supporting decision-makers or automated control mechanisms for example in autonomous weapon systems. In this paper we introduced a basic formalism allowing for dealing with uncertainty, and we discussed two approaches: a simple, naive method and the method based on Bayesian networks. The former requires a relatively small number of data, but it does not take into consideration a dependencies and a complex relations between anticipated events, which may result in inadequate probabilities. The latter is much more complex, because it requires declaration of not only a number of conditional probabilities, but also predicting various scenarios describing the results of decisions. Although this approach is complex, the predictions and probabilities calculated can be much more adequate.

Our model presents how to combine reasoning with abstract values with probabilistic reasoning. We use probabilities to represent uncertainty, but it is an open question (and a topic for our future research) whether probability is the most adequate way of representing uncertainty [16]. Our model is rooted in research on the well-known concept of expected utility, but we adapted this method to the problem of multiple values instead one notion of utility.

We also presented a simple example as a *proof of concept* describing how the formalism can be used to model and test IHL compliance when deciding on the use of weapon systems in armed conflicts.

[12] There are, however, a number of doubts concerning uncertainty of *what* should be considered: see the discussion in Sect. 3.

In a broader perspective, our model can be seen as an attempt to combine qualitative reasoning (represented by model of legal reasoning, scenario construction, etc.) with a quantitative one (represented by a probabilities calculus) and a platform for hybrid knowledge- and data-driven systems, where causal relations, probabilities can be extracted from past cases. It is open for further discussion and future research whether our model can be incorporated into other mechanisms, like Markov Decision Process (MDP) or Partially Observable Markov Decision Process (POMDP). However, unlike MDP or POMDP, our model is not created to introduce the decision making mechanism, but to set the limitations on the available decision set (to exclude decisions which lead to forbidden results).

It is also important to emphasize that our model was created on the basis of NATO regulations concerning the targeting process, giving this model practical application. It allows not only for the selection of legally admissible decisions, for doing this it in a transparent way, which is important from a legal point of view (see [17] for a detailed discussion).

7 Conclusions and Future Work

Autonomous military devices are currently the object of extensive discussion. Our model can be seen as a step into better understanding the nature of the military decision making process and as introducing a discussion on the possibility of creating IHL compliant military autonomous devices.

Although we focused on military devices, the problem we discuss here is much broader. Our work can be seen as an opening of a discussion how to reason with values in an uncertain environment. Is it possible to explicitly represent predictions and its relations to values? Can such models be helpful in decision making?

Our research also opens up other questions. Expanding qualitative reasoning with uncertainty, rather than pushing uncertainty to the boundaries of systems using qualitative reasoning, typically using thresholds to turn uncertain data elements into boolean propositions, will result in a larger decision space. This allows us to calculate the pros and cons of those decisions, and might make us choose a less likely good decision if the likely negative (side) effects outweigh the positive ones. This however is more computationally demanding and exploration of these operational consequences as well as the related sustainability issues that come with it are future research issues. Another issue is time critical applications, and we may assume that using these algorithms in automated devices may have to quickly produce good enough results. The meta-reflection on these meta-economical aspects of automated reasoning is to be addressed in future research as well.

We believe that by introducing a transparent model of the targeting process which encompasses a mechanism for dealing with uncertainty, we can help to make the military operations less harmful to civilians. It is also important to note that we are not advocating that such systems should be used without human

supervision: our position is instead that this complex topic requires rational debate, and that this paper is one contribution to such debate. It is then up to commanders, policymakers and legal experts to determine *whether* and *under what circumstances* such technologies may be used on the battlefield.

References

1. Prosecutor v. Gotovina et al., IT-06-90-T, Prosecution's Public Redacted Final Trial Brief, 2 August 2010, para. 549 (2010)
2. Office of the General Counsel, U.S. Dep. of Defense Law of War Manual (2015)
3. Proportionality in the conduct of hostilities: the incidental harm side of the assessment. Chatham House (2018)
4. 36, A., PAX: Areas of Harm - Understanding Explosive Weapons with Wide Area Effects. Article 36/PAX, Netherlands (2016)
5. 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)
6. Atkinson, K., Bench-Capon, T.: States, goals and values: revisiting practical reasoning. In: Proceedings of 11th International Workshop on Argumentation in Multi-Agent Systems (2014)
7. Atkinson, K., Bench-Capon, T.J.M.: States, goals and values: revisiting practical reasoning. Argument Comput. **7**, 135–154 (2016)
8. Bench-Capon, T.J.M.: Persuasion in practical argument using value-based argumentation frameworks. J. Log. Comput. **13**(3), 429–448 (2003)
9. van den Boogaard, J.: Proportionality in international humanitarian law. Ph.D. thesis, UvA-DARE (2019)
10. Cohen, A., Zlotogorski, D.: Proportionality in International Humanitarian Law: Consequences, Precautions, and Procedures. Oxford University Press, Oxford (2021)
11. Curtis E. Lemay Center: Air Force Doctrine Publication 3-60 - Targeting (2019). www.doctrine.af.mil/Doctrine-Publications/AFDP-3-60-Targeting
12. Ekelhof, M.A.: Lifting the fog of targeting: 'autonomous weapons' and human control through the lens of military targeting. Naval War Coll. Rev. **71**(3), 61–94 (2018)
13. 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)
14. Gisel, L.: The principle of proportionality in the rules governing the conduct of hostilities under international humanitarian law. ICRC International Expert Meeting 22–23 June 2016 (2016)
15. 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. Technical report, Geneva (2019)
16. Kahneman, D., Tversky, A.: Prospect theory: an analysis of decision under risk. Econometrica **47**(2), 263–291 (1979). http://www.jstor.org/stable/1914185
17. Kwik, J., Van Engers, T.: Algorithmic fog of war: when lack of transparency violates the law of armed conflict. J. Future Robot Life **2**(1–2), 43–66 (2021). https://doi.org/10.3233/FRL-200019

18. Kwik, J., Zurek, T., van Engers, T.: Designing international humanitarian law into military autonomous devices (2022). https://ssrn.com/abstract=4109286
19. North Atlantic Treaty Organisation: Allied Joint Doctrine for Joint Targeting, Edition A Version 1. AJP-3.9 (2016)
20. Oeter, S.: Specifying the proportionality test and the standard of due precaution: problems of prognostic assessment in determining the meaning of "may be expected" and "anticipated". In: Kreß, C., Lawless, R. (eds.) Necessity and Proportionality in International Peace and Security Law, pp. 343–366. Oxford University Press, Oxford (2020). https://doi.org/10.1093/oso/9780197537374.003.0012. https://academic.oup.com/book/33456/chapter/287736397
21. 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)
22. Russell, S.: Human Compatible: Artificial Intelligence and the Problem of Control. Penguin Publishing Group (2019). https://books.google.pl/books?id=M1eFDwAAQBAJ
23. Russell, S.: AI weapons: Russia's war in Ukraine shows why the world must enact a ban. Nature **614**(7949), 620–623 (2023). https://doi.org/10.1038/d41586-023-00511-5. https://www.nature.com/articles/d41586-023-00511-5
24. Schmitt, M.N.: Wired warfare: computer network attack and jus in bello. Int. Rev. Red Cross **84**(846), 365–399 (2002)
25. Schmitt, M.N.: Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations, 2nd edn. Cambridge University Press, Cambridge (2017)
26. Schmitt, M.N. (ed.): Tallinn Manual 2.0 on the International Law Applicable to Cyber Warfare, 2nd edn. Cambridge University Press, Cambridge (2017)
27. Schmitt, M.N., Schauss, M.: Uncertainty in the law of targeting: towards a cognitive framework. Harv. Natl. Secur. J. **10**, 148–194 (2019)
28. Schmitt, M.N., Thurnher, J.S.: "Out of the loop": autonomous weapon systems and the law of armed conflict. Harv. Law Sch. Natl. Secur. J. **4**, 231–281 (2013)
29. Shulsky, A.N., Schmitt, G.J.: Silent Warfare: Understanding the World of Intelligence, 3rd edn. Brassey's Inc., Washington, D.C. (2002)
30. Urbaniak, R., Kowalewska, A., Janda, P., Dziurosz-Serafinowicz, P.: Decision-theoretic and risk-based approaches to naked statistical evidence: some consequences and challenges. Law Probab. Risk **19**(1), 67–83 (2020). https://doi.org/10.1093/lpr/mgaa001
31. Verheij, B.: Proof with and without probabilities - correct evidential reasoning with presumptive arguments, coherent hypotheses and degrees of uncertainty. Artif. Intell. Law **25**(1), 127–154 (2017). https://doi.org/10.1007/s10506-017-9199-4
32. Vlek, C., Prakken, H., Renooij, S., Verheij, B.: A method for explaining Bayesian networks for legal evidence with scenarios. Artif. Intell. Law **24**(3), 285–324 (2016). https://doi.org/10.1007/s10506-016-9183-4
33. van der Weide, T.L., Dignum, F., Meyer, J.-J.C., Prakken, H., Vreeswijk, G.A.W.: Practical reasoning using values. In: McBurney, P., Rahwan, I., Parsons, S., Maudet, N. (eds.) ArgMAS 2009. LNCS (LNAI), vol. 6057, pp. 79–93. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12805-9_5
34. Winikoff, M., Sidorenko, G., Dignum, V., Dignum, F.: Why bad coffee? Explaining BDI agent behaviour with valuings. Artif. Intell. **300**, 103554 (2021). https://doi.org/10.1016/j.artint.2021.103554

35. Zurek, T., Kwik, J., Mohajeriparizi, M., van Engers, T.: Können autonome waffensysteme humanitäres völkerrecht anwenden? Tagesspiegel (Cybersecurity section) (2022). https://background.tagesspiegel.de/cybersecurity/koennen-autonome-waffensysteme-humanitaeres-voelkerrecht-anwenden. Accessed 20 Dec 2020
36. Zurek, T.: Goals, values, and reasoning. Expert Syst. Appl. **71**, 442–456 (2017). https://doi.org/10.1016/j.eswa.2016.11.008
37. Zurek, T., Kwik, J., van Engers, T.M.: Model of a military autonomous device following international humanitarian law. Ethics Inf. Technol. **25**(1), 15 (2023). https://doi.org/10.1007/s10676-023-09682-1
38. 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. Frontiers in Artificial Intelligence and Applications, Saarbrücken, Germany, 14–16 December 2022, vol. 362, pp. 273–278. IOS Press (2022). https://doi.org/10.3233/FAIA220479
39. Zurek, T., Mokkas, M.: Value-based reasoning in autonomous agents. Int. J. Comput. Intell. Syst. **14**, 896–921 (2021). https://doi.org/10.2991/ijcis.d.210203.001
40. Zurek, T., Wyner, A.: Towards a formal framework for motivated argumentation and the roots of conflict. In: Grasso, F., Green, N.L., Schneider, J., Wells, S. (eds.) Proceedings of the 22nd Workshop on Computational Models of Natural Argument, CMNA@COMMA 2022. CEUR Workshop Proceedings, Cardiff, Wales, 12 September 2022, vol. 3205, pp. 39–50. CEUR-WS.org (2022). http://ceur-ws.org/Vol-3205/paper5.pdf