The risks of autonomous weapons

A view from the South

Contributions to the debate on the use of Artificial Intelligence for armament purposes

M. Vanina Martínez
PhD in Computer Science

Ricardo O. Rodríguez
PhD in Computer Science

Since the beginning of history, the world has faced three main problems: famine, plague and war. Over time, humanity has organized and developed tools and processes to overcome them. In this sense, Artificial Intelligence (AI) is one of the latest inventions that like fire, the wheel, steam engines or atomic energy, will have a disruptive impact on the future of society.

Nowadays, AI plays an important role in the so-called Precision Agriculture where crops are controlled remotely. Also, it has played an important role in the control of the covid-19 pandemic, either by apps that track infections or the development of vaccines. Finally, AI is the promoter of a new arms race that needs to be controlled.

In this document we intend to reflect what we considered to be central technical issues for the debate on the control of lethal autonomous weapons (LAWs). We are aware that in the current state of the AI knowledge frontier, our opinions may generate more doubts than certainties. In this sense we will try to honor J.L Borges' maxim: "Doubt is one of the names of intelligence" .

The ultimate purpose of this debate is the creation of a legal framework that on the one hand, prohibits the use of fully autonomous weapons systems and on the other, regulates and controls the development of autonomous weapons systems that have meaningful human control. In doing so, we aim to avoid a dehumanized future, where armed systems can decide to kill and use force without a military authority understanding or being fully responsible for the consequences of such action.

A bit of context

Technical.

In a quick overview, we could say that algorithms are the implementation of concepts or models.

For example, there is the concept of "arithmetic sum" and there are different ways to implement it. On the contrary, most machine learning algorithms (MLA) can be broadly described as :

“An algorithm that seeks to approximate a concept/model through a limited amount (that can be very large though) of data/examples of it”

In this sense, MLA are meta-algorithms, that is, they implement the implicit concept/model given by data/examples. In this way, an MLA would obtain a conceptualization of the arithmetic sum from many sum examples (i.e., two summands and their result). The underlying assumption is that if we give a MLA enough examples, it will get a procedure that is equal to our known school method of addition.

In general, people do not learn how to operate numbers with numbers or examples. However, we learn to speak with examples. A MLA finds patterns and regularities in the set of examples and it generates a model based on them. It is important to note that large part of the scientific findings of the 17th century were produced by observation.

Nowadays, virtual models generated by MLA, depending on the app's domain, typically perform better than physical ones . In fact, there are several modelling tasks where MLA produce better results than a human expert. An example of this is the detection of tumours in tomography [8].

Essentially, the main challenges of this technology involve overcoming three types of problems. The first type is about the quality and quantity of needed data (labelling and bias). The second one is about the process of obtaining the model (overfitting, training vs testing, optimization, translation, etc). The last one is based on the obtained model (interpretability, predictability, explainability, transparency, accountability, etc).

Accuracy and efficiency are not the central weaknesses of MLA because both are increasing rapidly. On the contrary, we prefer to point out predictability and understanding as the core challenges of MLAs. The first is limited by a theoretical aspect [1]. The second has a section dedicated to it.

For the purpose of this presentation we will divide AI into two groups: algorithms based on knowledge (KBA) and those ones based on data (MLA). The latter present the most succesful technique nowadays and our analysis will be focused on them.

Political-diplomatic

In the political-diplomatic field there are many precedents that should be taken into account when putting in context the current debate about lethal autonomous weapons systems (LAWS). In the first place, the four major (bilateral and multilateral) arms control treaties: the Anti-Ballistic Missile Treaty (ABM), the Intermediate Range Nuclear Forces Treaty (INF), the Treaty on Open Skies and the Strategic Arms Reduction Treaties (STARTs); all tend to reduce East-West tension. Unfortunately, they have concluded or will not be ratified, leading to a very uncertain future of armed conflicts. Secondly, more auspicious precedents can be cited, such as the Treaty on the Prohibition on Nuclear Weapons (which entered into force on 22/01/2021) and the Conventions on Biological and Chemical Weapons, Cluster Munitions, Antipersonnel Mines and Blinding Lasers. All of these precedents, with its advances and setbacks, mark the international effort to regulate the development, production, commercialization and use of different types of weapons. On this basis, discussions on LAWS are being held.

In general, State parties to the Convention on Certain Conventional Weapons (CCW) agree that "[I]dentifying and reaching a common understanding among High Contracting Parties on the concepts and characteristics of lethal autonomous weapons systems could aid further consideration of the aspects related to emerging technologies in the area of LAWS" (2019 report 19 b). To this end, State parties continue to seek further clarification and a shared understanding of the specific technical characteristics that constitute or define "autonomy" or autonomous capabilities of "decision-making" in weapons systems. Until now, States have indicated that certain capabilities, including "autonomous functions in the identification, selection or engagement of a target" are "one of the essential characteristics of weapons systems based on merging technologies in the area of lethal autonomous weapons systems" (2018 report 19 a.). But they have not reached a shared understanding of a technical definition of autonomy and have highlighted "self-adaption; predictability; explainability; reliability; ability to be subject to intervention; ability to redefine or modify objectives or goals or otherwise adapt to the environment; and ability to self-initiate" as technical characteristics of

autonomous weapons that "may benefit from additional clarification or review" (2019 report, 20 a., b).

Nowadays, in context of the CCW there is a Group of Governmental Experts that received a mandate to "explore and agree on possible recommendations on options related to emerging technologies in the area of LAWS, in the context of the objectives and purposes of the Convention" *(see https://dig.watch/process/gge-laws)*. In this space of reflection and discussion, different groups with opposing interests are represented: International Organizations, NGOs, Governments, Armed Forces, the arms industry and scientists. It is precisely our intention here to contribute with solid elements from the academic sector to this debate.

In this space of reflection and discussion, different groups with opposing interests are being represented

In this sense we believe that technically such a debate should not only be circumscribed to respect International Humanitarian Law-IHL, since this restricts its analysis to the context of the battle-field in regular conflicts, but does not consider other extremely important aspects such as its design, manufacture and commercialization.

On the other hand, we believe that in relation to IHL, the major challenge lies in ensuring the principle of *accountability*. Other issues, such as the principle of *proportionality*, *opportunity*, *precaution* or *distinction*, which are very important to assess a military action, are potentially well parametrizable in a LAWS from a technological point of view. Although, we recognize that these principles require complex assessments based on the prevailing circumstances at the decision moment of the attack and also during one, we consider that they can be captured by logical rules and metrics of adequate performance.

Finally, we believe that we must be very careful to avoid irrelevant technicalities. For example, as it has been already demonstrated in the case of antipersonnel mines, it is essential to define precisely the object of prohibition/regulation. But given the complexity of the type of systems we are analyzing, it seems inappropriate to focus that definition in whether the system uses sensors to determine where and when the force will be used (see [7]). We believe that this is not a good discriminant and does not align with the complexity of weapons systems based in AI, both existing and to be developed in the near future. The first problem with this characterization is that in general, sensors are in the physical place where the action is deployed, but the decision can be made from a long distance from the place of execution, from data that does not directly come from the sensors, but from preprocessing and sometimes, even historical data that may not even be related to the entries of the system sensors. Therefore, in the case of LAWS, we believe that it is best to define them in a multi-purpose way and perhaps rather in conceptual terms as: their learning, perception, decision making and executing capabilities, which can then be carried out physically or digitally in different ways [10]. For example, by

learning through historical data, sensors could capture the current situation, interpretation module and decision making based on rules and physical actuators.

Contributions

Our contributions to the LAWS debate in this document will focus on three areas that we consider most relevant and urgent in the discussion:

- Meaningful human control.
- Predictability and reliability.
- Understandability, explainability and interpretability.

Before we begin to describe these three issues we are going to raise some concerns about automatic decision-making systems. A very important point about the nature of these concerns is that they can not be necessarily solved with an improvement in the qualitative and quantitative accuracy of MLAs. A first important issue that responds to the very essence of these systems is that they identify correlation and not causality, in other words, they discover "symptoms" but not "motives". This leads to a loss of reliability and the possibility of unpredictable behavior. It also questions the fact that automatized decisions diminish the possibility of legal and ethical objections to the results. In [7] the challenge that LAWS impose to human values is emphasized, and dehumanization and lack of understanding on how systems work is added to the above. Consequently, they may result in a failure to establish meaningful human control over risks to peace and safety, due to the fact that autonomy and remote activation "facilitate" taking the risk; on the one hand, as a result of political thresholds being significantly reduced and on the other, as a consequence of promoting automatized responses that escalate rapidly to an arms race.

Axe 1: Meaningful human control

Since the beginning of the negotiations mentioned above in 2014, States have proposed that the focus of the debate should be on the issue of human control, and on how it should be implemented on weapons, their critical functions, attacks, processes of target selection and (final) decisions on the use of force, etc. Although, most States agree that human control must be more meaningful than only the possibility of aborting an attack at the last moment, there is no such unanimity at the determination of how the human role should be defined and applied in the use of (lethal) force. In this section, we will attempt to order the main points of discussion and outline some methodological proposals. In order to do this, it is necessary to introduce some preliminary concepts.

In general, three levels of command can be recognized: (see [2]):

- **Strategic command**, which translates political purposes into military objectives.
- **Operational command**, which turns general objectives of the strategic level into specific tasks for the tactical forces.
- **Tactical command**, which directs the specific use of military forces in operations to execute the tasks given by the operational command. The tactical command covers the deployment of units, platforms, personnel not belonging to a formed unit and weapons systems that may be in direct contact with the parties of a conflict. In [3], this level is called "Mission Command".

MLAs can be used, in the first place, throughout the entire decision-making process of a military action. Nowadays, high and middle ranks use AI tools for data analysis and decision making. However, we believe that at the strategic and operational levels, there is no real possibility that, in the short and medium term, MLAs will make decisions with a level of autonomy that violates human dignity. For this reason is that we focus in what is referred as its tactical use. On the other hand, it is important to note that even military leadership seems reluctant to delegate its powers (see [3]) to a higher rank than mission command.

Also, it is very important to identify the four types of control that are often called into question. They are the following: (see [2]) :

- **"Human full control"**, where the system makes no decisions of its own, but it is remotely controlled.
- **"Human in the loop"**, where the system implements the ordered task with autonomy, but requires human ***intervention*** to validate and implement actions.
- **"Human on the loop**", where the system implements the ordered task with autonomy under the ***supervision*** of human operators who can correct or abort an specific action if necessary.
- **"Human off the loop"**, where the system implements the ordered task without any supervision or human intervention.

The loop usually refers to the sequence of tasks (some of which may not be on a specific action) in which the execution is usually divided:

- **Search**: finding the target, collecting information and performing intelligence on the battle field.

- **Location**: detecting and confirming the location accurately and estimating the time available.

- **Follow-up**: maintaining positive identification and updating information about the target and its environment.

- **Check-up**: assessing the rules of engagement (ROE), collateral damage and risks to the own forces.

- **Confrontation**: executing the attack on the basis of authorized, restricted and prohibited target sets with suspension and cancellation capability.

- **Evaluation**: assessing the effectiveness of the attack and determining whether a new attack is necessary and under what conditions.

With the introduction of all these elements we are able to develop the main issues of Meaningful Human Control.

The central issue is to ensure that human beings are ultimately responsible for the consequences of a military operation, even though the time and space distance between the operation directive is far from its execution in the battle field. This distancing and the unpredictability of the consequences it brings, in turn, raises concerns on the application of international human law, ethical acceptability and operational effectiveness. That is why we seek to define, what type and degree of human control is required in practice, in order to ensure that human beings continue to play their necessary role in decisions that involve the use of force in specific attacks in armed conflicts, while complying with legal, ethical and operational requirements, regardless of the technology's sophistication. All parties involved in the discussion recognize that human beings should have some type of control over weapons and the use of force in specific attacks during armed conflicts. Where questions such as *how* and *when* diverge, human beings should exercise such control in operational contexts. As highlighted in the GGE, there are likely to be some control measures that can be applied in all circumstances and others that would only be necessary depending on the context.

In [4] three types of control measures are identified:

- On the parameters for the use of weapons systems, including measures that restrict the target type and task for which a LAWS is used; placing time and space limits in its operation; restricting LAWS effects; and allowing fail-safe deactivation mechanisms.

- About the environment that controls or structures the use of LAWS (for example, use them only in environments where there are no civilians or civilian objects, or exclude their presence during the operations' duration).

- On human-machine interaction, with measures that allow the user to monitor the LAWS and intervene in its operation when necessary.

These control measures can help to reduce, or at least offset, the inherent unpredictability in the use of LAWS and mitigate the risks, especially for civilians. From a legal perspective, a LAWS operator must have enough control in order to have reasonable certainty about the effects of its use in an attack and be able to limit them according to what IHL requires. Certainly, ethical considerations can demand additional restrictions, especially in the use of force against civilians. In any case, the implementation of these measures should respect the balance between ensuring legal compliance, ethical acceptability and operational utility.

In the section 2 of [4], there are three difficulties presented by the LAWS in complying with IHL: number, context and unpredictability. Nowadays, value judgements that verify compliance, which also reflect ethical considerations, are part of the regular training of the armed forces. We understand that most questions about the qualitative or evaluative judgments that must be made in real time, in a complex and dynamic environment, to ensure respect for the principles of distinction, proportionality and IHL precautions, are technologically surmountable. Moreover, we think that defining standards and "parametrizations" on the subjective assessment of the safeguard of these principles, can be useful to rethink them. In this sense, we consider that the professional practice of a military could, at least partially, be emulated by an AI system in the not too distant future. To the extent of being able to pass an adapted Turing test. Current AI technology does not allow it yet but, from our understanding, it is only a matter of time. For this reason it is very important to define design criteria, validation, predictability and understandability metrics, and finally continuous auditing mechanisms, so that when the technology exists, it can be reliable and predictable, but above all, measurable and auditable. Certainly this is a true challenge and it will take a very important effort, but it is technically possible to overcome the three difficulties mentioned before. This does not mean that for political reasons, it is decided to preserve a relevant portion of control to human beings. Again, in our view, the decision is political and should not be based on legal or technical issues that do not go to the heart of the matter.

In order to establish our position with respect to this, we should say that in the subsection of [4], “Application of IHL in practice: what requirements for human control?”, two positions that emerged from the workshop of experts in June 2019 are presented. Technologists consider that the only limit for autonomy is technology: the more "sophisticated" a LAWS is, more tasks can be assigned to it and less user control is needed during an attack. Humanists, regardless of the LAWS technical features and IHL rules about hostilities conduction, demand assessments based in context and values from the people that plan, decide and execute attacks.

We consider that this position is manichean. The fact that LAWS can acquire increasingly "sophisticated" features does not justify granting them full autonomy. But also, the argument that IHL intrinsically requires people to have absolute control, is not completely correct. Again, in our view, the decision is political and the legal

or technical issues that do not go to the bottom of the matter, should not be taken into account. We agree with the authors of [4] that both approaches can be made compatible in all three practical control measures mentioned above, in a simultaneous and incremental approach of them all. By this we mean that they should be taken together, coordinating and expanding its scope. For example, if only the attack to military targets is parametrized, it should also establish the action's allowed context, such as duration and localization, and levels of communication/human-machine intervention. And all of this depending on the stage of the military action that is in process. It is not expected to require the same level of control during the search, location and follow-up, than during the confrontation itself. Also, it would be reasonable not to allow a LAWS to define the overall objectives of a mission at an early stage.

We will not comment on ethical issues in this document. Not because we do not find them important, but because ethical and moral issues about this domain have not reached enough maturity as to relieve it. What we can say is that most of the work that has ventured into this topic, assume a deontologist or regulatory approach. In this sense, we claim that the approaches to the ethical issues that arise from this philosophical school are liable to join a LAWS. Again, the limiting factor is not technological, but from another nature, political before, philosophical in this case.

In order to end this section, we would like to reproduce the following chart from [4] that summarizes the control aspects that could be integrated into the life cycle of a LAWS, as we have discussed previously:

Exercising what measures and when?

	Control over parameters	Control over environment	Control through human machine interaction
In use		• Complex understanding of the context	• Guarantee supervision • Intervention and deactivation capability
In deployment	• Set time and space parameters • Conditions to the application of force	• Set objectives and people limits • Set time limits and exclusion areas	• Explanation mechanisms • Include supervision, intervention and deactivation mechanisms
In design	• Objectives type and profile limits • Spatial and temporal controls	• Warning mechanisms • Safety and inviolable mechanisms	• Ensure mechanism for human supervision, intervention and deactivation • Train users

Axe 2: Predictability and reliability

As environments and ways of applying AI techniques increase, discussions about the effects that the use of AI systems may have over our society expand. As we mentioned previously, many of these critics are based on the very nature of AI systems that are used today, that is, those that use ML that would enable, in principle, the possibility of unpredictable behavior . In intuitive terms, predictability is the extent to which a system's effects can be anticipated. In [4] predictability is identified as an indispensable factor to comply with IHL norms, in particular the prohibition of indiscriminate attacks, the proportionality principle and the requirement of taking precautions in attacks, since operators must be able to limit the effects of the weapons they use.

[5] identifies three aspects of this principle that we will discuss below to understand the concept in depth and its relation to other criteria and concepts that we will discuss in this article.

From a technical point of view, predictability is seen as the ability of a system to execute a task with the same performance as in previous tests or applications. For ML systems this implies that the system has to be able to function in the same way as in its training stage. In computer systems this is directly related to their correctness.

The correctness property has to do with the system delivering the expected result given the initial configuration for a specific task. For example, if the system's task is to differentiate people from other objects that are presented to it through a video stream or a photo sequence, a possible correctness measure (which is established at the validation stage) is to calculate the rate of false positives and/or false negatives, that is, of all the people that were shown, how many it identified correctly as people, and how many of the objects that were shown were identified (mistakenly) respectively. This is only one of the many quality and performance metrics that are commonly used for machine learning algorithms. Determining which is the most adequate metric, depends on the specific problem to be solved; for example, a high rate of false positives, can not be tolerable for an application that tries to identify NNs in a database of missing people, but is crucial for a LAWS or a system that identifies criminals or suspects.

Therefore, from the point of view of software engineering, predictability is thought as a function that takes into account the system's correctness, i.e. the level with which the system can replicate and reproduce the same correctness over time, and the extent to which the system can adapt to allow the processing of different data to which it was exposed during the training and validation, without loosing performance.

---

1 *https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234*
https://medium.com/@MohammedS/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b

The latter is particularly important for machine learning systems, because an inadequate selection of the training and validation set can cause the system "memorizing" instead of "learning, meaning that it can not generalize the patterns found in the data that slightly diverge from the instances already seen. This situation is called overfitting in statistical terms.

On the other hand, from an operational point of view, predictability aims to have the possibility of anticipating particular actions of the system running. Any autonomous system is expected to have some degree of associated unpredictability, especially when the task is more complex and the context in which it operates is more dynamic, because it is impossible to anticipate every possible situation that the system will face once implemented. To better illustrate this aspect of predictability, we should go back to the system that we mentioned previously, which should differentiate people and objects from a video feed. Let's suppose that this video comes from a camera attached to the system that captures live images from the context where the system is moving. Regardless of the AI model used in such a system, it is not possible to determine in advance, in the system's design, every possible (type of) objects and people (and external events) that it can face when deployed in a physical environment, unless the environment is completely controlled or counts with (almost) perfect information[2]. In real life environments, where the information is imperfect and incomplete, operational unpredictability is a complex problem because it is not only difficult to anticipate what the system will find, but it can also be very difficult to anticipate how it is going to react to those events.

There are still no robust tools that identify and represent adequately relevant variables

In systems whose operation depends on the extraction of patterns from training data, this can be even more worrying. On one hand, it has already been demonstrated that in practice this systems can fail in a very unpredictable way when the data entries vary (even slightly) from the expected, because the way in which they fulfill their goals does not necessarily follow logic or reasonable patterns like a human being [11]. On the other hand, in order to be able to quantify how predictable these systems are, the quality of the data in relation to the deployment environment should be adequately quantifiable. For this, there are still no robust tools that identify and represent adequately (potentially) relevant variables for this task.

Finally, from a more global perspective, predictability, understood as the degree in which a

2 *https://es.wikipedia.org/wiki/Información_perfecta*

system's results and effects can be anticipated to its use, is determined by a lot of different factors that also have an impact in the technical and operational aspect. Among the most noteworthy factors for this article, as we have already mentioned, are the issues of data availability and quality (both for training and validation), the specific type of task or function that the system must solve, and the interaction with other systems (not only computing ones) in a dynamic and complex environment.

Predictability is a necessary factor, although depending on the domain, it may not be enough (in the next section we will analyze other complementary properties), to achieve a reliability relation with autonomous systems while they are running In fact, depending on the factors mentioned in the previous paragraph, being able to measure the level of predictability can be vital in the success of deploying systems that require human-machine interaction for the decision making. Specifically regarding autonomous weapons, predictability is extremely important for the effective exercise of different types of control that we discussed in the first part of this article. The capacity of the operator to assess how an autonomous system would respond in a given circumstance is absolutely necessary for the system to be able to define whether (and at what risk) it is feasible for the system to accomplish the task in a way that respects the different rules of engagement, IHL, etc., as well as the mode of operation required to make it so (configuration of specific parameters for the mission, for example).

It is interesting to note that for a system to be predictable, it is not necessary that the user knows or understands its operation (which is complementary, as we will discuss in the next section), predictability should be determined on the basis of the performance analysis of the system. The models that nowadays supports AI based systems present, as we have discussed, a lot of difficulties to assess its predictability in a trustworthy way. Nowadays large investigations are being carried out to improve its robustness.

These systems also impose a new challenge in the definition of design processes and their validation or testing. Even though there is a large variety of methodologies and tools in software engineering for traditional computer systems, to assure certain level of predictability, robustness and reliability, in most cases, they can not be applied directly. For example, the area of study of formal methods focuses on the development of mathematical techniques to describe both software and hardware systems and their requirements (what those systems are expected to do) so that what the implemented system does can be tested with mathematical operations which check if the system does what it is supposed to.

---

3 According to the UE, the document "ETHIC GUIDELINES FOR TRUSTWORTHY AI" states that "trustworthy AI has three components, which should be met throughout the system's entire life cycle: 1) it should be lawful, complying with all applicable laws and regulations; 2) it should be ethical, ensuring adherence to ethical principles and values and, 3) it should be robust, both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm". URL: Ethics guidelines for trustworthy AI | Shaping Europe's digital future (europa.eu)

These techniques are successfully used in general for closed systems of controlled complexity and high risk. However, they have serious efficiency problems, as the systems become more complex; beyond this, they do not directly apply to ML systems and it is not clear that something similar can be effectively achieved in the short and medium term . On the other hand, nor the plethora or testing tools that exist for traditional systems are adequately generalized for AI based systems. One of the most important problems is to be able to generate relevant test cases for the system's behavior outside the scope of design, in order to reproduce the environment it will find once deployed as reliably as possible, especially if you expect the system to function for example, in a physical and open environment (as would clearly happen for an autonomous weapon). Simulation tools can play a very important role in this because it is not necessary, in principle, to face the system with a real situation like a battle field. However, the problem of being able to predict unforeseen situations still remains, because the environment must be simulated in a computer system itself.

Axe 3: Understandability, explainability and interpretability

Other properties, complementary to predictability, have been identified as important for generating reliability in a AI based systems. In this section we will analyze some of the ones that are related to the human capacity of understanding how the sublaying model works or understands the reasons why the system delivers certain result or decision making, as well as the system's ability of providing relevant information about its operation or its decisions. In different proposals, these properties are grouped under the umbrella of "Transparency", that include the access to not only the process that the AI system carries out but also to all the processes that compose the system's life cycle from its conception and design to its development, deployment and evolution over time [6].

A system's understandability or interpretability is focused in understanding the AI model, that is to be able to identify why the system does what it does when it does. The spectrum of AI based systems vary widely as to how understandable and interpretable their fundamentals are. Analyzed from a strict sense, this property can require that a human user (developer or operator) can understand the complexities of the AI model. In this sense, a rule-based expert system, or a decision tree, relatively small, can be highly understandable for its designer, although a user can require some level of training to understand the logic of a knowledge structure and the inference process. On the other hand, the learning process in a neural (deep) network includes the assignment of millions of weights to features as the same identical network of input data (for example, the pixels of a photo or a video).

---

4 https://sites.google.com/site/sistemasexpertosunah/home/sistemas-expertos-basados-en-reglas

5 https://es.wikipedia.org/wiki/%C3%81rbol_de_decisi%C3%B3n

Although the designer/developer understands the meta process that is happening (the network is looking to optimize a mathematical function by adjusting those weights adequately), it is impossible that its mind can fully understand the calculation itself. For the system user, it may be that even the "intuitive" description of what it means to adjust a function is incomprehensible. The term "black box" is generally used to refer to models like the last one, where it is not possible to access the laws that govern the processing that happens inside of it.

What is the ignorance limit we can accept in these cases where sensitive issues, such as human life, depend on the system's operation?

If we take the term in a more general sense, certain degree of understanding or interpretability of an AI model can be achieved from an extensive observation of the system's behavior . Most human beings use smartphones in a very effective way without knowing exactly how the device works internally, but we construct a mental model of its operation, from continuous use and tutorials (specific training), in which we can trust, and only in a few occasions the system surprises us behaving in a completely unpredictable way. However we can question if achieving a degree of empirical comprehension by observing the system work is or not enough to manipulate any AI based system. Clearly in order to manipulate a smartphone that detects our face to unlock itself or that indicates what road to take towards an unknown location is enough, but what happens with an intelligent weapon? What is the ignorance limit we can accept in these cases where sensitive issues, such as human life, depend on the systems' operation?

In relation to the predictability principle, understandability is complementary and both are necessary the more complex the task that the system must solve and the environment in which the action takes place is. On one hand, a high degree of understanding of the system increases its predictability. However, predictability is not enough in itself, even when it exists in a high degree, especially to monitor if the system is working well or in those (maybe scarce) situations where the system fails in an unpredictable way. As we established in the first section, meaningful human control is key in the use of lethal autonomous weapons. Let's imagine the case where the system that has been deployed and operates together with a human operator, fails in an unexpected way; it is the human operator who must execute some type of control over the device in order to correct or avoid not wanted collateral damage. If the user as well as having been trained in the use of the system, understands at any level how its "logic" works, he could in principle, understand the situation quickly and carry out some contingency action. For example, the simple realization that the system is "not viewing or identifying" an object that the operator does notice, because it has never been fed with similar images, gives the operator tools to correct the situation or take the necessary precautions to fulfill the task. The combination of a high degree of understanding and predictability are critical for the effective use of an AI system that involves high risk for the user or the environment in which it is deployed.

Finally, we will address the explainability principle. This concept is one of the most developed ones in relation to AI systems, in fact, the XAI is an area of research in expansion within the AI field. Explainability is the AI based system's capacity of justifying its results or decisions in terms that a human being can understand. In order to understand the concept it is necessary to differentiate it from understanding and interpretability. A black-box system could, in principle, give explanations about its results, but this explanations would be as opaque and incomprehensible as the rest of the model. Even so, it could serve to generate user confidence if the same appeal to an affordable correlation of it. On the other hand, most of the AI based systems that we use nowadays, even those that are completely interpretable, are not designed to offer explanations that accompany its results or suggestions. Explainability requires a specific ability from the system, that is, it is the system who acts to give explanations, while interpretability leaves the system in a passive mode and it is the human who analyses it.

It is clear that if a system can explain itself, it can generate more confidence in the user than one that does not offer this function. However, there is a fervent discussion over if explainability is absolutely necessary (or even wanted) for any AI based system. In relation to intelligent weapons, an explicable system can be very useful for the operator's training stage, because it gives him the possibility of creating a more robust mental model of the system's behavior when it justifies its actions. On the other hand, in execution once deployed, this capacity can make human control more effective and fluid, especially when the system does not work as expected and it is difficult for the operator to understand if its working well or if it is failing.

It is important to note that the explainability concept is complex and determining what a "good" explainability means depends directly on the domain of application, the specific task that the system is solving and the (type) of user or operator. [13].

To conclude this section we want to note that, in the case of autonomous weapons, there are different positions over which of these properties are absolutely necessary, enough or even, useful. One of the purposes of this article is to make clear that predictability, understandability and explainability are important and a high-risk system like an intelligent weapon, can not afford to do without any of them. With the three, and by being thought from the system's design and for the rest of its life cycle, they can improve the chances that an operator can interact effectively with that system, exercising the necessary control over it and mitigating the possible failure risks or collateral effects that may happen. The degree to which these properties are present depend on each type of system expected to be developed.

Proposals

The authors of this document are convinced that the international community has been accumulating a lot of effort to agree on criteria to preserve humanity's interests in different areas, and especially in medicine. We consider that the ethical and technical criteria used in the health area can serve as a guide to set guidelines in the context where AI is used in arms. In particular, in recent years, the International Medical Device Regulators Forum (IMDRF) has set important definitions to the regulation of the use of AI in medical devices. Something that emerges quickly from reading the existing documentation, is that regulatory models are still under development[6]. This is essentially due to the fact that the scientific community has not still given a technical-formal response with respects to how to measure predictability and reliability or how to interpret or explain the emerging behavior in an AI system generated with examples. It is important to be clear that this technology is at an early stage of progress and we are making conjectures about its effects based on speculation, not evidence. In this sense, we need to be cautious with our projections without keeping from setting ethical and social limits to the impact of these technologies. The problem that we face is that in order to set limits we need to have metrics. For example, for a drug to be commercialized there are protocols that take years of experimentation to verify its safety and effects. But we are far of something similar for AI systems. For example, we talk about bias, where we have examples that we denounce, but we do not have a precise specification of the concept, nor a way to measure it and much less a way to mitigate it. The same is reproduced for various concepts we have addressed in previous sections. All, while new ML architectures (with new challenges) keep being proposed.

The above aims to support the idea that the control and regulation of LAWS should be thought as a continuous process in time that should generate preventive limits that evolve and adjust with the correct advances in technology. A clear example of this is tools of facial recognition. Its abusive use in crime prevention has promoted precautionary actions against its use. So, companies, like Amazon, Google and Microsoft decide to suspend sales of systems that use this technology. On the other hand, States establish temporary moratoriums or prohibitions with the intention of reaching more specific regulations. This does not mean that facial recognition technology will never be used, but its use is conditioned to social agreements and specific evidence that make its use safe.

Following this dynamic, we are going to divide our proposals in two levels: a more general (and political) one, and other more specific (and technical) level.

General proposals:

- The establishment of a cyber ethical code of good professional practices that LAWS' designers

---

6 See for example: https://algorithmwatch.org/en/story/medical-devices/

and developers should adhere to in the same way that they sign a confidentiality agreement. It could also serve to establish a hippocratic oath of computer knowledge. These values should be incorporated into the curricula of both military training and computer professionals.

- The formation of an Agency for Regulation and International Control of LAWS. This agency would be dedicated to generate validation standards and metrics, perform audits and monitoring, and analyzing situations in which regulations fail in order to feed back into the regulation and control system.

- Encourage the establishment of LAWS free zones, following the current idea of areas free of nuclear weapons.

- Encourage the formation of an international registry of LAWS manufacturers and prohibit the manufacture of LAWS without license (see [12]).

- Enabling companies that design and produce LAWS to have an internal ethics committee whose members have stability and freedom to carry out the task. That would solve the issue of business confidentiality.

Specific proposals:

- We believe that we should change the focus from regulation and talk about: “Software as a military device” (MD), as any piece of software intended to be used for one or more military purposes that perform without being part of a hardware device. On this basis, it is possible to extend the concept to MD based in Artificial Intelligence and Machine Learning. From there it is possible to set regulations that give place to safe operations for civilians that respect IHL.

- To achieve LAWS reliability and verifiability objectives, it is necessary to provide them with “black boxes” (such as the ones used in planes) that allow to assess its posterior actions.

- Development of a segmentation or the different types of LAWS that allow to establish specific legislation for each one. The first try carried out in [7] is interesting, especially if we consider the issue of "tying" the division on the basis of the existence of sensors. As we mentioned before, we believe that the segmentation must be based in multi-purpose aspects, that include both the device's features and capabilities, and configurations of the performance environment and the target tasks. In [14] various segmenting criteria appear, that we believe to be a good starting point.

- Promote technological research that led to better LAWS assessment in at least the three analyzed axes.

References

- [1] Learnability can be undecidable. Shai Ben-David, Pavel Hrubeš, Shay Moran, Amir Shpilka and Amir Yehudayoff. Nature Machine Intelligence, VOL 1, pp. 44-48. JANUARY 2019.
- [2] The Human Element In Decisions About The Use Of Force. Infographic from INIDIR. Merel Ekelhof y Giacomo Persi Paoli.
- [3] Mission Command and Armed Robotic Systems Command and Control A Human and Machine Assessment. Robert J. Bunker. Land Warfare Paper 132 / May 2020. The Association of the United States Army.
- [4] Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control. Vincent Boulanin, Meil Davison, Netta Goussac y Moa Peldán Carlsson. June 2020. SIPRI.
- [5] The Black Box, Unlocked: Predictability and Understandability in Military AI. Holland Michel, Arthur. 2020. Geneva, Switzerland: United Nations Institute for Disarmament Research. doi: 10.37559/SecTec/20/AI1
- [6] First draft of the Recommendation on the Ethics of Artificial Intelligence (UNESCO 2020): https://es.unesco.org/artificial-intelligence/ethics
- [7] Regulating Autonomy in Weapons Systems. https://article36.org/wp-content/uploads/2020/10/Regulating-autonomy-leaflet.pdf
- [8] Artificial intelligence is improving the detection of lung cancer. Elizabeth Svoboda Nature. November 2020. https://www.nature.com/articles/d41586-020-03157-9
- [9] Interpretable machine learning. A Guide for Making Black Box Models Explainable. Molnar, Christoph. 2019. https://christophm.github.io/interpretable-ml-book/.
- [10] Artificial intelligence: a modern approach. 3rd ed. Russell, S. J., Norvig, P., & Davis, E. Upper Saddle River, NJ: Prentice Hall. 2010.
- [11] Robust Physical-World Attacks on Deep Learning Visual Classification. K. Eykholt et al., 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, pp. 1625-1634. 2018.
- [12] Model Law against the Illicit Manufacturing of and Trafficking in Firearms, Their Parts and Components and Ammunition. United Nations document. 2011. https://www.unodc.org/documents/legal-tools/Model_Law_Firearms_Final.pdf
- [13] Explanation in artificial intelligence: Insights from the social sciences. Miller, T. Artificial Intelligence (2019).
- [14] A choices framework for the responsible use of AI. Benjamins, R. AI Ethics 1, 49–53 (2021). https://doi.org/10.1007/s43681-020-00012-5

Maria Vanina Martinez PhD in Computer Science (University of Maryland, USA) with a post-doctoral degree in Oxford University. Researcher at the Institute for Computer Science (CONICET-UBA) in the area of Artificial Intelligence and professor of the Computer Science Department, in the School of Exact and Natural Sciences at University of Buenos Aires, where she teaches the subject “Ethics & AI”. Member of the National Committee of Ethics in Science and Technology of the Ministry of Science, Technology and Productive Innovation. Member of the Stop Killer Robots campaign and the Sehlac network.

Ricardo Oscar Rodriguez PhD in Computer Science with a major in Artificial Intelligence. Associate professor in the Department of Computing, FCEyN-UBA and member of Institute for Research in Computer Science (UBA-CONICET). His research is based in the development of logical models for reasoning under incompleteness and uncertainty. Currently, teaches the subject “Ethics & AI" and is member of the Sehlac network and the Stop Killer Robots campaign in which he actively participates. Ha has been co-chair & financial chair of IJCAI2015 in Buenos Aires.


APP
Asociación para Políticas Públicas


CAMPAIGN TO STOP
KILLER ROBOTS


SEHLAC
SEGURIDAD HUMANA
EN LATINOAMÉRICA Y EL CARIBE