Towards Broadening the Perspective on Lethal Autonomous Weapon Systems' Ethics and Regulations

Bianca Ximenes[2], *Diego Salcedo*[3], *and Geber Ramalho*[4]
Federal University of Pernambuco

1. Introduction

LAWS may, without the explicit approval of a human being, decide to cause harm to or kill people. Their adoption involves complex ethical, technical, commercial, legal, regulatory, strategic, and geopolitical issues. That is why, in the scope of IHL, taking into account the CCW, a United Nations GGE has been created to debate the governance of this emerging technology.

2 Informatics Center.
3 Information Science Department.
4 Informatics Center.

The two opposing positions on LAWS could be total banishment or no regulation. Without making any judgment of the value of these two positions, both supported by some countries so far, we prefer to adopt an in-between perspective, since intermediary solutions raise more interesting and complex debate than adopting one of the two positions. Indeed, supposing that these kinds of weapons could be authorized or adopted, some questions should be answered: in which cases can they be adopted? Under which circumstances can they be used? Which kinds of weapons can be fully automatized? How does one limit the damage of the use of such weapons? Who is accountable for their use? What are the adoption criteria and processes for these weapons?

Several advances have been made on LAWS governance by the GGE/LAWS, establishing the basis and premises that may enable an agreement or convention on the topic. In particular, the CCW/GGE has, in their late 2019 session, converged to form the 11 Guiding Principles for LAWS, representing an excellent starting point for more detailed discussion (CCW/GGE.1/2019/3).

Our reflections on LAWS issues are the result of the work of our research group on AI and ethics at the Informatics Center in partnership with the Information Science Department, both from the Federal University of Pernambuco, Brazil. In particular, our propositions and provocations are tied to Bianca Ximenes's ongoing doctoral thesis, advised by Prof. Geber Ramalho, from the area of computer science, and co-advised by Prof. Diego Salcedo, from the humanities. Our research group is interested in answering two tricky questions: What would an ethical AI be? And how can one guarantee that a given intelligent system will follow intercultural human ethical principles?

In this paper, we explore these two questions in two sections, in the hope of slightly broadening the perspective of the current

LAWS debate. In section 2, we show that there are discussions and research works currently being conducted worldwide on ethics and AI in general, which could shed a light on the particular debate on AI and weapons. Indeed, the task of ethics is to determine the elements that allow us to have and build intercultural dialogue. In section 3, we draw attention to the various forms of regulation beyond law or any kind of formal mechanism such as conventions. LAWS involve such complex issues, with critical consequences on humanity, that the debate and the solutions should not neglect all possible kinds of regulations.

2. Ethics for Artificial Intelligent Systems and how it affects LAWS

The more AI adoption advances in society, bringing socio-economic benefits, the more ethical questions are posed to governments, companies, and citizens on topics such as employment (certain human occupations will disappear, while new vacancies will be created); privacy (citizens leave digital tracks, but have little control over this data); and automation of decisions (which may be unfair and/or incomprehensible). On the latter, the most promising machine learning techniques, such as deep neural networks, involve complex models that cannot explain their decisions in a way that is understandable to the citizen. In addition, algorithms can incorporate bias against certain groups, as it is exemplified in the famous case of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system, which tended to lengthen prison sentences for black people in the United States (Kirkpatrick, 2017; Spielkamp, 2017). Therefore, discussing the application of ethics in AI is becoming a hot topic in universities, enterprises, and governments.

2.1. Ethics and AI

From a practical dimension of the debate, ethics is not synonymous with morality. Ethics alludes to the collective, morality is about the behavior of an individual. Ethics, therefore, lends itself as a justification for the daily practices of people and organizations. If, on the one hand, ethics is, in philosophy, one of the three major fields of study, along with epistemology and metaphysics, on the other hand, it is a practice of uninterrupted reflection on choices, behaviors and decisions with the constant objective of the improvement of social life. Ethics is the collective debate in search of the corporate model that we, at present, want for our future, in this sense, it is a defense of intelligence, our dialogical and decision-making condition for the coexistence of the collective, the community, the groups, and that, to this day, persists in our socio-cultural practices, precisely in moments of greatest intellectual challenge.

Therefore, to discuss what an ethical AI would be, it is worth recognizing that ethics is a human concern and pursuit. Machines, even the ones presently considered intelligent, are far behind human "generalist intelligence". They do not comprehend the context of which they are a part. The IEEE (Institute of Electrical and Electronics Engineers) Ethically Aligned Design Manual (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019) warns about how misleading it may be to attribute to autonomous systems an anthropomorphic intelligence they do not possess.

Concerning this matter, Loh summarizes (Loh, 2019):

> It is currently assumed that technological developments are radically changing our understanding of the concept of and the possibilities of ascribing responsibility. The assumption of a transformation of responsibility is fed on the one hand by the fundamental upheavals in the nature of "the" human being, which are attributed to

> the development of autonomous, self-learning robots. On the other hand, one speaks of radical paradigm shifts and a corresponding transformation of our understanding of responsibility in the organizational forms of our social, political, and economic systems due to the challenges posed by robotization, automation, digitization, and industry 4.0. It is also expressed widely that, thanks to these circumstances, our modern mechanized mass society sets ultimate limits to responsibility, even opening up dangerous gaps in the possibilities of attributing responsibility.

The discussion on how to translate the principles of an intercultural human ethics to a machine, or to AI, is complex for many reasons: privacy concerns, responsibility for autonomous action, delegation of decision making, transparency, bias in collected and analyzed data, surveillance, and AI opacity. Isaac Asimov, in the 1950s, had already established firmly that robots, machines, and every other possible kind of artificial intelligence might be ***logical,* but not *reasonable***. And the inherent pondering that ethics brings about has to do with reasonability more than logic, as slight differences in context bring about completely different preferences and results. A good illustration as an example is the trolley problem and all of its posterior adaptations. (Ahlenius & Tannsjö, 2012; Goldhill, 2018; Judith Jarvis, 2008; Thomson, 1976; Waldmann & Dieterich, 2016; Ximenes, 2018)

2.2. Floridi's principles

Artificial Intelligence Ethics discussions have reached international spheres, and they are mapped in the AI Ethics Guidelines Global Inventory[5]. Another initiative worth mentioning

5 Available at <https://www.rri-tools.eu/-/ai-Ethics-guidelines-global-inventory>.

is the Algorithm Watch[6], an organization committed to evaluating and shedding light on the algorithmic decision-making processes that have compiled most of the AI Ethics manuals proposed so far.

In the current profusion of Ethics manuals and guidelines for AI, Prof. Floridi's AI4People framework emerges as the foundation for any serious discussion on the subject. (Floridi et al., 2018; Floridi & Cowls, 2019)

In this work, inspired by Bioethics principles, Floridi and colleagues from the Digital Ethics Lab at Oxford University propose five overarching principles highlighted as being the most important to be taken into account:

- **Beneficence** refers to a practice where the priority should maximize the benefit and minimize the loss. It may also be understood as promoting overall well-being, preserving dignity, and sustaining the planet. In some sense, institutions and states that have AI will be in a great position to create value if AI is used as a means to improve beneficence rather than diminish the well-being of citizens. "The prominence of beneficence firmly underlines the central importance of promoting the well-being of people and the planet with AI." (Floridi & Cowls, 2019, p. 4)
- **Non-maleficence** highlights precisely the main characteristic of the principle of beneficence. Thus, it establishes that the action must cause the least damage (action that does not do harm). In this sense we could propose, as examples, problems related to privacy, security, and misuse prevention for avoiding doing harm while trying to do good. As Floridi and Cowls comment, "it is not entirely clear whether it is the

6 Available at <https://algorithmwatch.org/en/>.

people developing AI, or the technology itself, which should be encouraged not to do harm." (Floridi & Cowls, 2019, p. 5)

- **Justice** establishes equity as a fundamental condition; thus, it is an ethical value in which each individual (agent) must be treated in accordance with what is morally correct and adequate and given what is due. The main characteristic of this principle is impartiality: acting with others disregarding their social, cultural, religious, financial and distinct aspects that may interfere negatively in the relationship. As put by Floridi and Cowls, "the diverse ways in which justice is characterized hints at a broader lack of clarity over AI as a human-made reservoir of 'smart agency'." (Floridi & Cowls, 2019, p. 6)
- **Autonomy** requires agents to have the skills and competencies to make decisions in a way that is respected for that. The vulnerability of agents, in specific or contingency circumstances, needs to be considered with respect to the decisions that will need to be made. In the sense of AI, Floridi and Cowls conclude that "the autonomy of humans should be promoted and that the autonomy of machines should be restricted and made intrinsically reversible, should human autonomy need to be protected or re-established."
- **Explicability (or Explainability)** is the need to understand and hold to account the decision-making processes of AI. This should be possible by providing intelligibility and responsibility to machine decisions through an accurate methodology in the core of the AI system that has implemented into itself a model of explicability. This is needed because there is a novel reality about AI: its functionalities and processes are invisible or unintelligible to almost all individuals. For Floridi and Cowls, this principle is possible, but also required, by

> "enabling the other principles through intelligibility and accountability." (Floridi & Cowls, 2019, p. 7)

Even though these principles seem abstract, requiring more precise guidelines for developers and decisors' daily activities, they do represent a good foundation for understanding what would constitute an ethical AI. This may be useful in the present context because LAWS are one of several specific applications of AI, and all such applications should ideally be adherent to the overarching principles of ethical AI. Beneficence, non-maleficence, and autonomy are more easily connected to the LAWS debate, and aspects related to each of these three tenets are mentioned throughout the 11 principles presented in the GGE/LAWS 2019 document (GGE LAWS, 2019). However, explicability is not explicitly mentioned in none such principles, and only (b), (d), and (h) are related to this vital aspect of building ethical AI through auditability, compliance, and accountability.

In traditional computer science, auditability has to do with the possibility of examining the source-code. However, machine learning, neural networks, and more modern and powerful AI techniques are black-box models by their very nature, making these systems harder to audit because they are not inherently explainable. The patterns found in data are often unclear to humans. It is also even more complex to determine accountability because part of the optimization and decisions is done according to parameters that AI engineers cannot directly control in detail. Therefore, extra effort has to be made by engineers and practitioners in order to provide clarity and explanations based on the model inputs and outputs. Research in the area of XAI (Explainable AI) is a growing field with more solutions and novel approaches being released every month. (Adadi & Berrada, 2018; Biran & Cotton, 2017; Gilpin et al., 2019; Powell et al., 2019) Typically, current solutions involve using statistical tools to probe for biases and building secondary computational or

mathematical models that approximate the system's behavior and optimization function.

There are different levels of explicability that can be provided, and they vary according to the criticality of the application and the level of expertise of the user. (Google, 2019) For instance, in a Netflix recommendation of movies to watch, a wrong recommendation does not carry heavy consequences, and the great majority of users are not specialists. Hence, the level of certainty of a recommendation is not as critical, and no explanation is given concerning how the AI system decided what to recommend. On the polar opposite, applications such as cancer diagnostic systems based on image detection carry heavy consequences for all people involved. Hence, they need further explanation to support the system output and diagnosis. LAWS are more similar to the second case, as they are employed in critical scenarios that have impact on life-or-death issues. Besides, explanation and auditing are normally carried out by experts in the area, who need more detailed information to make decisions or determine accountability. Users are not the only ones who benefit from explainability, as understanding the systems also carries benefits to legislators, legal departments, and engineers themselves, as it becomes possible to audit the models at some level, but this discussion is out of the scope for this paper.

Considering the benefits of having more explainable systems, we argue that a possible next step for the CCW/GGE Principles could be considering the need of providing some level of explainability and which the necessary metrics are to be used to determine whether a system should be deployed and used or not. The exact thresholds might be the subject of more debate, but it is important for practitioners creating these systems to understand the requirements, and that everyone involved understands what these autonomous systems take in as parameters to make decisions. Especially so because in some contexts (i.e. defending against attack), humans will be completely

out of the loop due to the need to respond promptly, and auditing and reviewing the decisions will be vital *post facto*.

2.3. Human role in machine-based decisions

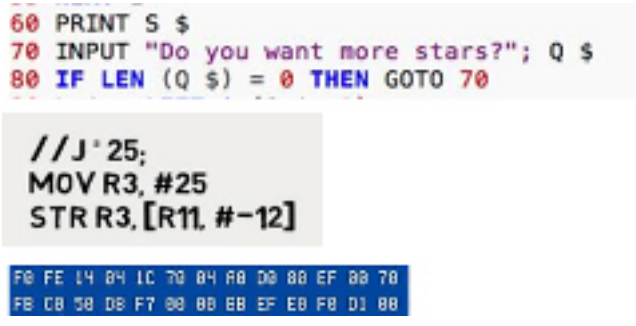
A central well-known issue in LAWS discussion is the role of humans in machine decision-making, commonly grouped into 3 categories, from high control to no one: Human in the Loop (HITL); Human on the Loop (HOTL); and Human out of the Loop (HOOTL). This discussion is, of course, not restricted to LAWS, even though, due to its criticality, it is especially applicable to this scenario. (Danks & Danks, 2013; Hoff & Bashir, 2015; Murphy & Woods, 2009)

In our research group, we have been trying to elicit some criteria currently used to decide the desired level of automation indecisions. We started by examining two domains, electricity distribution and Intensive Care Units, since these domains are highly regulated and involve risky decisions. We have found dozens of criteria. (Gilboy et al., 2011) For example, in US Emergency Rooms, these criteria are compiled in the ESI (Emergency Severity Index). Some of them may be useful or are already being used in the LAWS debate, such as: time to act (how much time is available for the decision); human factor (what the consequences of the decision on people's life are), environmental impact (what the consequences of the decision on the environment are); cost (what the overall cost of the automation is and what savings are generated by it); responsibility (how easy the identification of the responsibilities for the decision is); concurrency (how automation positions me in the face of competition), technical complexity (how complex and reliable the implementation of the automation is). The point here is to stress that it is important to establish a clear set of criteria on when to adopt fully automated decisions, how to do it, why do it, and who is able to do it. In the domains of electricity distribution and Intensive Care Units, this discussion seems to be more mature than in LAWS. We argue that, by

ensuring that every step taken in conceiving and building a system, from data collection to deployment, is ethical, we may equally ensure that a given intelligent system will have an overall ethical behavior as a consequence.

We can also explore the issue of computer autonomy from the engineering point of view. In computer science, we work with the notion of layers of abstraction. Each layer increases the level of abstraction, which means that, as we go to the upper ones, it is simpler to program a machine. The first layers are related to hardware, from the silicon substrate itself to the electronic boards. On top of that, there are the layers related to software, going from the machine code to Application Programming Interfaces, passing through assembly and programming languages. For those who are not familiar with these technicalities, imagine that, when one presses a brake or pushes the car's accelerator, this person does not need to know all the mechanical and electronic gears, mechanisms, and components involved in braking or accelerating the car. For the sake of clarity, this is an abstraction of the actual structure. Figure 1 illustrates three levels of software layers:

Figure 1: Example of three software layers. From the bottom, we have machine code, then assembly, then a simple programming language



What is AI from this point of view, after all? It is another layer of abstraction on top of programming. For instance, instead of programming the behavior of the machine, AI techniques allow the programmer to set only goals and rules, because the system has an embedded "inference engine" that knows how to start from a fact to deduce new facts according to the rules. Or the programmer can just give some examples of a given concept and let the machine learn the rules.

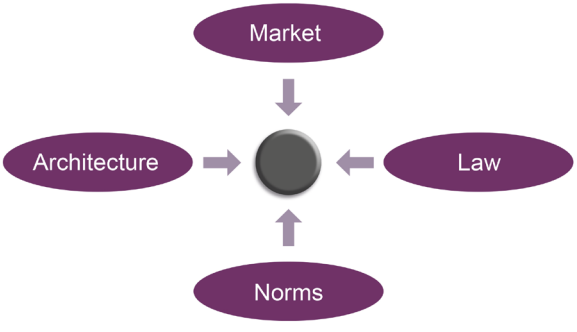
Thus, abstraction is necessary and the natural evolution of computer programming. The problem is that **the more abstract, the easier to program, but less control the programmer has over the machine**! So, the popular fear of losing control of machine decisions is not just a laic concern. Building ethical AI is a complex issue not only in philosophical terms, also fromthe technical point of view. (Russell, 2019)

3. Regulations for Artificial Intelligence and how it impacts LAWS

Half of the 11 principles proposed in the CCW/GGE 2019 document explicitly mention law, in particular compliance with International Humanitarian Law. There is no doubt that International Humanitarian Law is a fundamental reference to the debate on LAWS, and that it sets boundaries for what is allowed or prohibited. Moreover, the reference of this type of law in the context of a diplomatic debate is even more natural. However, if in the previous section we tried to broaden the perspective of the LAWS debate by pointing out that there are more AI artifacts than weapons in the ethical debate, in this section we want to emphasize that there are more forms of regulating AI artifacts than only laws. This is especially truer in the age we live, when technology is ubiquitous, and communication barriers have decreased, enabling a plethora of possible regulations.

Indeed, regulation is a broad concept. And in this context, we perceive regulation as any force or influence that changes the behavior of an agent, being able to limit or otherwise modify its actions. In order to establish our context and examples, we have adopted the Pathetic Dot framework proposed by Lawrence Lessig, which is very useful inexplaining and systematizing the discussion about regulation forces in the Internet era. (Lessig, 2006) Lessig explains that, from the point of view of someone or something that is being regulated, this entity is constrained by the inter-relations of four main forces, which are always balanced. Those forces are norms, laws, market, and architecture. The interaction between those forces can strengthen or undermine the influences of one upon the others, and their action is dynamic, changing across time. Figure 2 below illustrates the Pathetic Dot framework.

Figure 2: Pathetic Dot framework describing each regulating force



The specificities of each force are briefly explained below. We emphasize that the reader should keep in mind that many instruments and tools of regulation are an amalgam of different types and generate an influence in more than one sphere.

3.1. Laws

These are the most formal types of regulation, represented by constitutions, statutes, and legal codes. (Lessig, 2006) Laws are able to formally regulate and enforce not only the Pathetic Dot itself, but other forces as well. It can counteract norms or reinforce them with legal resources, limit market liberties, and define ideal architectures. While laws can be a highly effective form of regulation, in a democracy where representatives are elected, they also depend on several participants to write, redact, and push them forward from proposal to actual piece of legislation. Besides, laws are based on behavior that has already occurred, which by nature carries the consequence that law is implemented after it is necessary, as it is defined *post facto*. It does not have the intention or ability of foresight, and the phenomenon that it regulates must be well-described and understood. Code-based systems change too rapidly for lawmakers to describe and understand the phenomena they create in a timely manner. Therefore, even though the agents' compliance is supervised, and the law's punitive power exceeds that of any other means as well, it is important to realize the relevance of other regulatory forces, even if only as a means to compensate the inherent delay in the creation of applicable legislation. That is why, in this paper, we urge participants of the LAWS debate to open their minds to other possibilities of regulations, which could perhaps be as effective as formal laws.

3.2. Norms

Norms are essentially social constructs. They reflect relationships, culture, and behaviors of a given community. Norms are often informal and might never reach a written format, being instead based on the notion of what is acceptable or customary to do, then being an example of what should be done, in a cycle. Even so, some behaviors and habits can be recognized as especially desirable or in

need of standardization and may be registered in different ways. For instance, books on etiquette attempt to systematize norms. Best practice manuals for code maintainability and readability are sets of norms. ISO/IEC certifications are a way of auditing and asserting that certain norms are being followed, and they are valuable because society places value in such certifications. They differ from licenses, for instance, because they are not mandatory or enforced by the state; they are created by communities and maintained by private companies or the third sector. Norms are enforced by social pressure, and not complying with such terms might lead to loss of social capital and graver consequences, such as ostracism. In the examples presented above, none of the norms *must* be observed, but they might bear social consequences. For instance, not observing etiquette might mean not being invited for another dinner in the future; not complying with code maintainability and readability practices might mean losing the job; not having an ISO/IEC certification might mean losing clients to another company that has it.

Norms underlie all social relations and are not always explicit. An example of implicit norms would be Google's Project Maven for LAWS, which caused developers and AI engineers from Google to resign and walk out of their offices as a means to oppose the company's decision to participate in military projects. This happened because workers did not have the same expectations and moral code as the company, hence the fallout. This brought social and market repercussions to Google and spurred them to discontinue military collaborations. (Shane et al., 2018; Statt & Vincent, 2018)

Our research group elaborated, as a reference, a proposition of an Ethical AI Certification for companies in the private sector, based on the Great Place to Work and B Corporation certificates. Such certifications are recognized by society as a seal endorsing specific behaviors and qualities. This means they communicate value and are able to calibrate trust and expectations about a given product,

service, or company. We opted for a practical approach, and cross-referenced the five AI principles discussed by Floridi and introduced in the previous section with an extended CRISP-DM framework, one of the most popular Data Science frameworks, which describes a pipeline for creating data-based products covering elements from understanding business objectives to data preparation, to model training and deployment. (Wirth & Ripp, 2000) This produced a matrix, in which we are considering what a company should do to address, such as the issue of explicability in the data preparation phase. Answers are posteriorly audited against company evidence and depending on how the questions were answered. The company is awarded the Certification (which was dubbed CEIA – *Certificação em Ética para Inteligência Artificial*, in Portuguese; translated as Certification in Ethics for Artificial Intelligence). Some examples of the 48 questions from the reference questionnaire we are proposing are listed below:

- *Are the impacts of the positive effects of your system mapped in a clear and accessible way for all the company through the business targets and quarterly goals?*
- *Is it possible for humans to review and change decisions made by AI systems developed in the company that are used in critical settings?*
- *In data acquisition and preparation, is there a company-wide guideline for the target population to be represented equally, avoiding inherent data bias?*

These are just some examples, but they translate the ethical position of the company concerning its AI applications and the maturity of discussions and actions taken in relation to its positions. The CEIA then assumes a two-pronged approach: it guides processes internally, while communicating company priorities to employees and employers, and it also communicates to the outside world (e.g.: clients, citizens, third parties) what to expect from that company's

AI products and services, which can correspond to several economic advantages that are further discussed in the following subsection, "Market," the third regulating force we will explore.

3.3. Market

Markets are where economic exchanges take place. Simplistically, supply and demand curves meet at a marketplace and undergo adjustments to reach an equilibrium, which defines the existent quantity of a specific good or service, and the price at which it will be sold. Market equilibrium is dynamic, and these changes allow for market regulation of entities. Supply and demand curves may suffer shocks and be displaced, achieving new equilibria. Agents may also deliberately change their propension to buy or sell, also achieving new equilibria outside efficiency allocations in their original curves.

We can cite some examples of market regulation for AI. Buyers may boycott a company due to scandals, due to invasion of privacy or any kind of ethical issues. The lost reputation can be fatal for a company's survival. For instance, after Microsoft's facial recognition system was identified as being biased, they rapidly improved their training datasets and overall results. Even so, this piece of news harmed Microsoft's results in the quarter, and the company released a statement to investors explaining how biased or flawed systems can hurt the company's image, and why it is important to improve these models. (Gershgorn, 2019)

Another recent example, seen in the World Economic Forum 2020, is the decision of investment funds to condition their investments to projects that are committed to environmental, social and governance (ESG) issues. This positioning drives the market to a different direction. In the future, these premises may include Ethics for AI systems. Simultaneously, some government units stated that they will no longer purchase and deploy AI systems that cannot offer intelligible explanations for their decisions, in cases where

the decisions directly affect people's lives. For instance, New York City outlawed the use of black box models in the public sector, and Pennsylvania opted to have the state create its own recidivism risk assessment model, and have the code open for inspection, since it will not be proprietary to a private company. (Campolo et al., 2017; Pennsylvania Commission on Sentencing, 2019)

Insurance companies also play an important role in regulating markets. Suppose that the accuracy of weapons in distinguishing civilians from military targets is low. If a mistake is made, someone will have to pay a compensation for it, and the insurance company can be called upon to cover it. This will exert market pressure on the improvement of weapons accuracy, for instance. This will also exert pressure to create industrial benchmarks for LAWS, establishing quality standards for such systems.

3.4. Architecture

Architectural force has to do with the structure and the design of things, and how they can mold behaviors and regulate the way people operate. Unlike the other forces, architecture is an intrinsic aspect of the entity being regulated, a characteristic. An everyday example is airport benches and their armrests. These armrests are often static and cannot be elevated, which makes it more difficult, if not downright impossible, for a person to lie down and occupy multiple seats at once. This has to do with the bench's *architecture*, its structure. The act of lying down could be regulated in multiple ways, such as (a) by outlawing the act and arresting the person (which may sound preposterous in an airport setting, but it happens in park benches all over the world, where the homeless might be arrested for loitering); (b) by normatively embarrassing the person through insistent glares and disapproving looks or, (c) more lightly, by putting signs requesting that people think of other tired passengers and do not occupy more than one seat at once; or yet, (d) by applying

a monetary fine if the person is caught lying down, combining law and market forces to regulate chair occupancy in airports. All these have the same goal, but we consider that the architectural approach is more direct and more likely to work, not only for airport benches but also for AI systems. Indeed, some architectures can be easily changed (e.g. the items displayed on a software, such as the first screen of a mobile phone), and others are more permanent (e.g. the https protocol to safely transfer data packages online).

Changing the architecture implies changing what something is as well as how it should operate. It is deeply connected to the concept of feasibility and what a system consists of (i.e. code is the building brick of software). However, changes in architecture can also be applied to less concrete things, such as processes. Changing the steps of a pipeline generates structural changes and new demands not only throughout the process, but also in the final result. For instance, the inclusion of automated testing and quality assurance steps in software development and industrial pipelines spurred practical changes in tasks and processes, and had direct results on final systems and goods. Therefore, there are multiple ways to influence how things are through architectural changes.

Similarly to our certification proposal, our research group has also conceived a Consumer Artificial Intelligence Information Leaflet (CAII), similar to Patient Information Leaflets (PILs) that accompany medicines. Drugs have different compositions and purposes, but they have uniform processes, tests, and standards the pharmaceutical company must follow to get them approved. (US Food and Drug Administration, 2019) We find that this heavily resonates with AI issues, as we were able to draw a parallel with the drug approval process (based on material from the FDA-USA, TGA-Australia, and ANVISA-Brazil). In all cases, it is necessary to undergo four phases: application, clinical studies, approval process, and post-market tests. In the application phase, the company must

provide the basic data about purpose, application, dosage, and overall population characterization, which is highly applicable to AI products. The clinical studies encompass the effects of the drug on the human body, its efficacy, safety, and side effects. These can be adapted to the AI context and consider safety, security, biases, and a study of social impacts. In the approval process, for both cases, results are audited and checked for compliance with current applicable laws. Finally, in the post-market tests, the efficacy and consequences of the product are tested on a large scale. In the end, all the highlights are condensed and provided in a single leaflet that is freely circulated and to which everyone can have access. Even though this approach blurs the lines of individual regulation forces, covering laws, norms, and market, it also helps understand *how* to build an ethical AI system and guides architectural decisions and processes.

Architecture is the main factor we influence in computer science, and this is what we consider a key for building ethical machine learning systems, LAWS included. We believe the most critical changes and decisions for ethical systems must be made, by design (as stated in Principles (c) and (g) of the CCW/GGE 2019), before any final product exists. The most efficient regulation from the standpoint of a system is one that imposes constraints while the system is being created, as it limits from the very beginning what a system can or should do. Applicable product constraints will be elicited according to principles discussed in intercultural forums, such as the CCW/GGE forums.

Considering LAWS, we might take into account, during the construction of the system, specifications that comply with ethical standards. For instance, if one is creating land mines that should not be activated by a person, only war tanks, the weighing sensors used must be able to identify and differentiate weights, so the mine is only triggered in the correct context. These sensors must be embedded in the device during the process of its construction,

before the product is ready. Similarly, one might create a certainty threshold for image recognition, so that if an attack is conducted with LAWS and the target is unclear, the weapon remains locked and cannot take further action. For instance, it may be acceptable to proceed with 80% certainty of what the target is—military personnel or civilian, vehicles, buildings, among others, and this information is automatically calculated by any machine learning model. For systems with a human in the loop, this level of certainty (i.e. precision/recall) can be shown on screen so decisions are made with the correct information; for systems with humans out of the loop, this information can be automatically taken into consideration as a condition to initiate an attack.

The crucial question is *how* to do that, as the relevant aspects must be considered beforehand to create a system known as "ethical by design." This recognizes that observing ethical principles cannot be incidental, but instead must be planned and built into the system itself. Floridi himself explains that this ethical design is about an approach model that can protect and promote the aforementioned ethical tenets (specifically the AI decision-making processes), thus incorporating them from the beginning into the design specifications, functional and non-functional requirements of technologies (e.g.: AI, robots, etc.), procedures, practices or infrastructures. Our aim here is not to document every single ethical consideration for an AI project, but to consider and propose, as a debate that should be self-evident, that an AI project ought not to advance the proliferation of unchecked and unaccountable weapons. (Taddeo & Floridi, 2018)

4. Final remarks

In this article, we have discussed the worldwide debate that has been ongoing for a few years concerning what would be considered an ethical AI and how to achieve it. The core discussion is not about whether LAWS should be allowed or banned, but instead about

how they are part of a broader scenario of AI ethics and systems, and where to look in order to advance the discussion in practical ways. We have presented the five guiding principles for ethical AI, and argued that the pillar of explainability should be directly considered in the CCW/GGE Principles, since it allows for a better understanding of the AI systems, what they can do, how likely they are to achieve specific goals, as well as establishing the ground zero for any discussions on compliance, auditability, and accountability that are vital for LAWS. We also addressed the discussion on the role of human beings in machine decision automation, remembering that adopting AI techniques and tools simplifies programming, but also implies a certain loss of control.

We have also broadened the discussion on regulation under the lenses of the Pathetic Dot framework for the Internet era and code-based products. Laws are one of the four determining forces that regulate any entity, but it is possible to incentivize other behaviors and the production of accountable systems through normative, market, and architectural forces. We posit that architecture is a strong and often overlooked regulatory force as it depends on deep technical knowledge, but it also corresponds to reliable results in a myriad of scenarios beyond LAWS and beyond AI applications, as it shapes the very structure and capability of a system beforehand. We have illustrated our argument with some practical tools for regulating AI systems we have created in our research group on Ethics and AI; these were the Certification for Ethical AI and the Consumer AI Information Leaflet. Such tools could be applied to the LAWS scenario as well.

The debate on ideal LAWS and ethical artificial intelligence have much in common and would benefit from sharing more common ground. Furthermore, alongside the intercultural forums, we also need interdisciplinary forums where we can unite legislators, thinkers, practitioners, and idealists to define what to pursue for

the future of humanity alongside artificial intelligence. Only then will we be able to identify a wide range of approaches, opting for the more efficient ones that comply with our ethical principles and moral values.

References

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*(c), 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Ahlenius, H., & Tannsjö, T. (2012). Chinese and Westerners Respond Differently to the Trolley Dilemmas. *Journal of Cognition and Culture*, *12*(3–4), 195–201. https://philpapers.org/rec/AHLCAW

Biran, O., & Cotton, C. (2017). Explanation and Justification in Machine Learning: A Survey. *IJCAI Workshop on Explainable AI (XAI)*, *August*, 8–14.

Campolo, A., Sanfilippo, M., Whittaker, M., & Crawford, K. (2017). *AI Now 2017 Report*.

Danks, D., & Danks, J. H. (2013). THE MORAL PERMISSIBILITY OF AUTOMATED RESPONSES DURING CYBERWARFARE. *Journal of Military Ethics*, *12*(1), 18–33. https://doi.org/10.1080/15027570.2013.782637

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.8cd550d1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and

Recommendations. *Minds and Machines*, *28*, 689–707. https://doi.org/10.1007/s11023-018-9482-5

Gershgorn, D. (2019). *Microsoft warned investors that biased or flawed AI could hurt the company's image*. Quartz. https://qz.com/1542377/microsoft-warned-investors-that-biased-or-flawed-ai-could-hurt-the-companys-image/

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2019). Explaining explanations: An overview of interpretability of machine learning. *Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018*, 80-89. https://doi.org/10.1109/DSAA.2018.00018

Goldhill, O. (2018, February). Philosophers are building ethical algorithms to help control self-driving cars. *Quartz*. https://qz.com/1204395/self-driving-cars-trolley-problem-philosophers-are-building-ethical-algorithms-to-solve-the-problem/

Google. (2019). *People + AI Guidebook*. https://pair.withgoogle.com/

Group of Governmental Experts on Lethal Autonomous Weapons Systems (GGE LAWS). (n.d.). *Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems*. https://undocs.org/en/CCW/GGE.1/2019/3.%0A

Hoff, K. A., & Bashir, M. (2015). Trust in Automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Judith Jarvis, T. (2008). Turning the Trolley. *Philosophy & Public Affairs*, *36*(4), 359–374. https://doi.org/10.1111/j.1088-4963.2008.00144.x

Kirkpatrick, K. (2017). It's not the algorithm, it's the data. *Communications of the ACM*, *60*(2), 21–23. https://doi.org/10.1145/3022181

Lessig, L. (2006). *Code: And Other Laws of Cyberspace, Version 2.0* (2nd ed.). Basic Books. http://codev2.cc/download+remix/Lessig-Codev2.pdf

Loh. (2019). Responsibility and Robot Ethics: A Critical Overview. *Philosophies*, *4*(4), 58. https://doi.org/10.3390/philosophies4040058

Murphy, R. R., & Woods, D. D. (2009). Beyond Asimov: The Three Laws of Responsible Robotics. *IEEE Intelligent Systems*, *July/August*, 14–20.

Pennsylvania Commission on Sentencing. (2019). *Adopted Sentence Risk Assessment Instrument*. http://pcs.la.psu.edu/guidelines/adopted-sentence-risk-assessment-instrument

Powell, A., Joshi, A., Carfantan, P.-M., Bourke, G., Hutchinson, I., & Eichholzer, A. (2019). *Understanding and Explaining Automated Decisions*.

Russell, S. J. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*.

Shane, S., Metz, C., & Wakabayashi, D. (2018). How a Pentagon contract became an identity crisis for Google. *The New York Times*. https://www.nytimes.com/2018/05/30/technology/google-project-maven-pentagon.amp.html

Spielkamp, M. (2017). Inspecting algorithms for bias. *MIT Technology Review*. https://www.technologyreview.com/s/607955/inspecting-algorithms-for-bias/

Statt, N., & Vincent, J. (2018). Google pledges not to develop AI weapons, but says it will still work with the military. *The Verge*.

https://www.theverge.com/2018/6/7/17439310/google-ai-ethics-principles-warfare-weapons-military-project-maven

Taddeo, M., & Floridi, L. (2018). Regulate artificial intelligence to avert cyber arms race. *Nature*, *556*(7701), 296–298. https://doi.org/10.1038/d41586-018-04602-6

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2019). *Ethically Aligned Design A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems First Edition The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems*.

Thomson, J. J. (1976). Killing, Letting Die, and the Trolley Problem. *The Monist*, *59*(2), 204–217. https://doi.org/10.5840/monist197659224

US Food and Drug Administration. (2019). *Artificial Intelligence and Machine Learning in Software as a Medical Device*. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

Waldmann, M. R., & Dieterich, J. H. (2016). Throwing a Bomb on a Person Versus Throwing a Person on a Bomb: Intervention Myopia in Moral Intuitions. *https://Doi.Org/10.1111/j.1467-9280.2007.01884.X*, *18*(3), 247–253. https://doi.org/10.1111/j.1467-9280.2007.01884.x

Wirth, R., & Ripp, J. (2000). CRISP-DM : Towards a Standard Process Model for Data Mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, *24959*, 29–39. https://doi.org/10.1.1.198.5133

Ximenes, B. H. (2018). Non-intervention policy for autonomous cars in a trolley dilemma scenario. *AI Matters*, *4*(2), 33–36. https://doi.org/10.1145/3236644.3236654