Telecommunications Policy 44 (2020) 101953

Contents lists available at ScienceDirect

Telecommunications Policy

journal homepage: http://www.elsevier.com/locate/telpol

Ethics of autonomous weapons systems and its applicability to any AI systems

Ángel Gómez de Ágreda

PhD. Candidate Universidad Politécnica de Madrid (UPM), Spain

ARTICLE INFO

Keywords:
AI ethics
Meaningful human control
Autonomous weapons
Explainability
CCW
Dual-use AI

ABSTRACT

Most artificial intelligence technologies are dual-use. They are incorporated into both peaceful civilian applications and military weapons systems. Most of the existing codes of conduct and ethical principles on artificial intelligence address the former while largely ignoring the latter. But when these technologies are used to power systems specifically designed to cause harm, the question must be asked as to whether the ethics applied to military autonomous systems should also be taken into account for all artificial intelligence technologies susceptible of being used for those purposes. However, while a freeze in investigations is neither possible nor desirable, neither is the maintenance of the current status quo. Comparison between general-purpose ethical codes and military ones concludes that most ethical principles apply to human use of artificial intelligence systems as long as two characteristics are met: that the way algorithms work is understood and that humans retain enough control. In this way, human agency is fully preserved and moral responsibility is retained independently of the potential dual-use of artificial intelligence technology.

1. Introduction

Recent developments in artificial intelligence (AI) and global competition have increased the tempo of R&D in this field. Data collection to feed algorithms raises serious concerns about matters like privacy that are seldom assessed correctly and deeply enough. While most uses of these technologies benefit humankind, they are also prone to be utilised for nefarious purposes, whether by the dual-use application of the technologies or through the illicit employment of the data (Aicardi, 2018; Penney, McKune, Gill, & Deibert, 2018).

This duality –civilian and military, beneficial and aggressive– of the use of AI technologies makes it necessary to take into account the possibility that tools designed for good may be used as weapons and, therefore, to regulate their development from that perspective.

Most treatises on the art of warfare agree with Carl von Clausewitz that war is political in nature. Its goal is not so much the destruction of the adversary as the subordination of their will to our own. Therefore, any tool that may affect people's freedom should, from an ethical point of view, be treated as a weapon and be subject to international humanitarian law (IHL).[1] Freedom should not be less prominent in our priorities than life itself only because of "… anxiety about the loss of human control over weapon systems and the

E-mail address: angel@angelgomezdeagreda.es.

[1] The late Russian General-Major Ryabchuk went even further, noting that "thought is the first to join battle. Indeed, thought is a weapon". (V.D. Ryabchuk), "Problems of Military Science and Military Forecasting under Conditions of an Intellectual-Informational Confrontation", Military Thought, no. 5 (2008), pp. 67–76.

https://doi.org/10.1016/j.telpol.2020.101953
Received 24 June 2019; Received in revised form 15 March 2020; Accepted 16 March 2020
Available online 31 March 2020

use of force …" (International Committee of the Red Cross, 2018).

The possibility of the weaponization of digital tools makes it very relevant to examine how weapons systems are dealt with and whether their regulation should be applicable to all AI systems susceptible of being so used.

It would make little sense to apply different standards to AI systems specifically tailored not to comply with the first of Asimov's Laws of Robotics and for those AI systems that will serve the same function even if they were not primarily built with that purpose in mind.

Rapid advances in these technologies make it imperative and urgent to give more in-depth consideration of the ethical implications of their development. This is especially true and evident in kinetic[2] robotic applications, but it is no less important for other areas in which the use of data originating from or delivered to humans might affect their decision-making process.

Most general-purpose codes of conduct and principles on AI either completely neglect or give only marginal consideration to these other uses of technology. This paper examines twenty-four of the most relevant and important codes of conduct to extract their common features. Controllability and explainability turn out to be key issues that need to be taken into account when designing an AI system. Unintended and unforeseen uses of AI are far more relevant than their presence in most current codes indicates.

The paper examines the work of the United Nations Group of Governmental Experts (UN GGE) as a subsidiary body of the "Convention on Certain Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects" (CCW). This group meets regularly in Geneva to deal with ethics and the regulation of military uses of AI involving lethal autonomous weapons systems (LAWS), sometimes referred to as "killer robots" (Horowitz, 2016).

There is no agreed-upon definition of LAWS. Nations and institutions have issued a variety of definitional approaches (United Nations Institute for Disarmament Research, 2017). All of them, nonetheless, include the use of robotic systems with a varying degree of autonomy to identify, select and engage a target with lethal consequences. Understanding of the term "autonomy" and the centrality of lethality are still a matter of debate (Boddington, 2017). For the purposes of this paper, LAWS will refer to any weapon system that has the ability to perform those functions with limited human intervention.[3]

The recent publication of the "Recommendations on the Ethical Use of AI by the Department of Defense" (Defense Innovation Board, 2019a) and its supporting document (Defense Innovation Board, 2019b) shows how timely studies like this one are. Defence institutions and armed forces are the leading actors in both the development and the use of the hardest versions of AI, and their thoughts and criteria are therefore of paramount importance in these matters.

This paper begins by identifying the key principles present in most ethical codes. These principles are then used to compare the different codes among them and to determine how LAWS are dealt with in each of them. Weapons systems are, nonetheless, made up of a set of technologies, each of which could have dual (civil-military) use that needs to be taken into account. It is useful, then, to examine the ethical principles being discussed at the CCW to determine their applicability to the rest of AI systems.

2. Key principles of ethical codes

This paper examines the present situation of ethical codes regarding AI, be they from academia, the institutions or the corporations. The most relevant and cited codes have been selected, as well as the most recent ones. A complete comparison is not presented in this paper and only the results are extracted to draw conclusions from them. A more detailed comparison is available in the table in Appendix A.

This methodology has been used previously in other relevant papers, either in this field or in other sciences (Baura, 2006; Upshaw Downs & Swienton, 2012; Rothenberger, Fabian, & Arunov, 2019). The most recent and the largest so far is the study conducted by the Chinese Academy of Sciences and the China-UK Centre for AI Ethics and Governance (China-UK Research Centre for AI Ethics and Governance, 2018).

Six of the codes were generated by or for official institutions.[4] A recent relevant document which is also included is the one issued by the Defense Innovation Board of the Defense Department of the US (Defense Innovation Board, 2019a), which provides the

[2] Kinetic weapons are those which use the kinetic energy contained in an object to produce damage as opposed to, for example, electromagnetic weapons. Kinetic robotic applications are, therefore, those which act physically.

[3] The lack of an established meaning makes it difficult to even set the terms to discuss (Ekelhof, 2017). Terms and definitions of intelligence itself as a human specific feature (Muehlhauser & Helm, 2012) and the value of human autonomy are still in dispute. This paper intends to transcend geopolitical considerations and to provide arguments for those who design, develop and employ AI systems at large.

[4] UNESCO's Report of COMEST (World Commission on the Ethics of Scientific Knowledge and Technology) on Robotic Ethics, (COMEST, 2017), Defense Advanced Research Projects Agency (DARPA) (Leys, 2018), EU Parliament (European Parliament Committee on Legal Affairs, 2016), EU Commission (Craglia, 2018), the UK House of Lords (Select Committee on AI of the House of Lords, 2018) and the OECD Principles on AI (OECD, 2019b; 2019a).

perspective of the user rather than that of the developer. The following eleven codes come from the world of academia and think-tanks.[5] The remaining four were generated by the industry and reveal both the ethical standards and the vision of the companies.[6]

Most of the codes were drafted within the last three years in Western nations or corporations. A comparative study reveals a high level of coherence among them and a sufficient degree of maturity.[7] A common denominator can be inferred that identifies the most recurrent concepts and their synonyms. The high correlation among the authors and the tendency to draw principles from already published codes are among the reasons for this uniformity.[8]

The table in Appendix A identifies the key principles that constitute this common denominator. They mostly coincide in all but name with those acknowledged by the Chinese Academy of Sciences in its wider study of close to fifty sets of principles.[9] Section 3 briefly discusses these dimensions.

Only a marginal number of these codes take dual uses of AI into explicit consideration. A few of them acknowledge the existence of non-pacific uses but choose not to deal with them, even if they can hardly be disassociated. Only two codes, IEEE (EAD2) and UNESCO's Report of COMEST (World Commission on the Ethics of Scientific Knowledge and Technology), give in-depth attention to LAWS.

This paper continues with a brief discussion about warfare and the role of dual-use AI technologies in modern conflict. The political nature of war makes it a confrontation of wills (Sun Tzu, Clausewitz). Hybrid conflicts, in what the military call the "grey zone", make use of all sorts of tools as weapons.

The following chapter moves into the study of the papers and reports issued at the UN Convention of Certain Weapons during the last two years. Earlier documents have also been reviewed, but they do not incorporate significant differences from the ones cited here. While this is not the only forum in which autonomous weapons are being discussed at the intergovernmental level (OECD, 2018), the UN CCW is widely regarded as the main reference when it comes to LAWS, and the most inclusive one.[10]

A comparison of the key concepts derived from the twenty-four ethical codes studied and the Guiding Principles proposed by the CCW ensues. This will allow the determination of the commonalities between both sets of principles and the suitability of the latter as a baseline code of ethics for all AI.

The technology associated with AI will not disappear. It is being incorporated into our everyday reality until it becomes ubiquitous and often transparent or invisible to the user. Initiatives that seek its total or partial eradication (Campaign to Stop Killer Robots, 2018), and those that demand a moratorium on its research and development (Bolivarian Republic of Venezuela, 2018)[11] until an overall vision of its possible perverse effects is developed, lack feasibility. Modern societies are neither willing nor in a position to renounce the benefits derived from their implementation (Belgium, Ireland, & Luxembourg, 2019; China, 2018). Therefore, a realistic approach to the consequences of the development of these technologies needs to include their unintended uses.

The identification of these perverse effects is a prerequisite. This will help mitigate them without precluding the implementation of the positive ones. It will also assist in deciding the way in which the next steps are taken. Feasibility, not only in technical terms, but also in political terms, must be one of the preconditions for any ethical or legal code to be developed (CCW, 2018). Disregard for political and economic interests will no doubt render any proposal inapplicable.

This leads to the need to establish specific requirements for the development of technologies associated with artificial intelligence and the preservation of human values in the new mixed context of carbon-based and silicon-based intelligences. In this new scenario, if we want to preserve our human values, new tools and procedures will be needed. New ethical and legal codes need to be agreed upon.

Ethics do not apply to machines, as free will is a prerequisite for the development of ethical standards. Moral values need to guide

[5] AI4People (Floridi et al., 2018), IEEE's Ethically Aligned Design (EAD2), Future of Life's Asilomar Principles, (Future of Life Institute, 2017), The Forum on the Socially Responsible Development of Artificial Intelligence's Montreal Declaration (University of Montreal, 2018), The tenets of AI (The Partnership on AI, 2018), Engineering and Physical Sciences Research Council (EPSRC)'s Principles of Robotics, (Boden et al., 2010), Biocat's Barcelona Declaration for the Proper Development and Usage of Artificial Intelligence in Europe (Barcelona Bar Association, 2019), Fairness, Accountability and Transparency in Machine Learning (FAT/ML)'s Principles for Accountable Algorithms and a Social Impact Statement for Algorithm (Diakopoulos et al., 2018), UNI Global Union's Principles on Ethical AI (UNI Global Union, 2018), The Association for Computer Machinery (ACM)'s Code of Ethics and Professional Conduct (ACM, 2018), and the most recent by Beijing Academy of Artificial Intelligence (Beijing Academy of AI, 2019).

[6] Google (Pichay, 2018), IBM (IBM, 2018), Microsoft (Microsoft, 2019) and DeepMind (DeepMind, 2017).

[7] Additionally, Tencent's Pony Ma proposed an ethical framework for AI with four distinctive principles: Availability, Reliability, Comprehensibility and Controllability (ARCC), which basically mimics other codes (http://www.linking-ai-principles.org/term/407).

[8] At the time of writing, the World Economic Forum (WEF) has created six councils to provide guidance and governance of several topics, including AI. The members of the Global AI Council are also relevant figures who have already taken part in earlier initiatives, and the principles so far expressed under the auspices of the WEF do not diverge significantly from them. https://www.weforum.org/agenda/2019/05/these-rules-could-save-humanity-from-the-threat-of-rogue-ai/.

[9] "Linking Artificial Intelligence Principles, Keywords and Topics", Research Centre for Brain-inspired Intelligence, Institute of Automation, Chinese Academy of Sciences (http://bii.ia.ac.cn/); Innovation Academy of Artificial Intelligence, Chinese Academy of Sciences; School of Artificial Intelligence, University of Chinese Academy of Sciences; China-UK Research Centre for AI Ethics and Governance (https://ai-ethics-and-governance.institute/), 2018–19. http://www.linking-ai-principles.org/keywords.

[10] NATO has not developed any formal code of conduct on ethical guidelines for LAWS. It simply advises to be mindful of the need to reach definitions and maintain transparency throughout the negotiations that will ensue. It also expresses its concern with "placing the decision to take a human life in the hands of a machine" (Williams & Scharre, 2015).

[11] Motion presented on behalf of the Non-Aligned Movement.

technologists in their research and to help decision-makers regulate the use that is made of the possibilities that AI presents from their deeper understanding of the possibilities and the consequences. However, responsibility lies mostly with the users choosing to utilize technology in immoral ways. The goal is that they should be beneficial for humanity as a collective body and for humans individually (and as equitably as possible).

Many of the most recent studies on the effects of artificial intelligence are questioning not the autonomy of machines, but that of the humans, who design and use them. While there is some ongoing discussion about whether or not rights should be awarded to robots (Gunkel, 2018; Tavani, 2018), it is human autonomy that needs to be protected (COMEST, 2017; Floridi et al., 2018). The humanistic vision of technological development is instrumental when it comes to constructing the social and political environment that results from the use of robots.

3. Codes of conduct comparison

The last few years have witnessed the publication of a large number of codes of ethical principles on AI. Close examination of twenty-four of them (see table in Appendix A) enables us to reduce these to a smaller number of principles that are contained in the vast majority of them in one way or another:

3.1. *Beneficence*

The first idea that appears to be present in most of the aforementioned codes is that of beneficence. Algorithms need to be a force for good, to promote both social and individual well-being. Yet individual and group interests may not always be as closely aligned as these proposals suggest. There are conflicting rights that need to be balanced. Algorithms designed to be socially beneficial may end up being so only at the expense of the benefit of certain individuals.

3.2. *Human dignity*

Human dignity is a controversial term that plays a central role in most ethical codes. It can be defined as “the recognition that human beings possess a special value intrinsic to their humanity and, as such, are worthy of respect simply because they are human beings” (Trinity International University). Ethics as a whole revolves around this idea. All other principles make little sense without due regard to individual rights and the qualitative difference between humans and the rest.

3.3. *Privacy*

Most of the proposed ethical codes also include privacy as one of their principles. It goes way beyond the purpose of this paper to explore all the different derivatives of this topic. Arguably, privacy is the basis for the rest of the principles discussed here. If data are the building block of decision-making, then access to those data and the way they are stored and processed constitute the primary concern from which all the rest evolve. That is especially true when current technology allows for a multidimensional understanding of the subject and not just a specific part of it. However, there is little appetite from the public to preserve their data (Potoglou, Patil, Gijon, Palacios, & Feijoo, 2013), due to a poor understanding of the implications.

Privacy is often associated with identity, anonymisation and transparency. Privacy may be both individual and group-related. Establishing a univocal digital identity for persons or groups amounts to the digital equivalent to DNA editing.

3.4. *Human autonomy*

Human autonomy, already mentioned above, is also addressed in most codes. The debate turns to the questions of how much autonomy remains for humans faced with the growing presence of AI in human society, and how autonomous machines affect human rights.

Guaranteeing that humans retain agency –understood as the capacity, condition, or state of acting or of exerting power– revolves around the concept of freedom and independent thinking. Ultimate control over the AI system needs to remain with the human operator in each phase of the decision process. Yet, even when humans retain control over machines, other algorithms may still be influencing the operator’s choices, in the form of assistance systems that provide the background information for decision-making. These assistants are at the same time far less capable and yet more influential in our mental processes than most humans perceive them to be (Dahlmann & Dickow, 2019), resulting in overconfidence in their decisions (Logg et al., 2019).

3.5. *Fairness*

Justice and fairness are two concepts that appear mostly in relation to unbiased decisions by algorithms. Biased decisions taken by algorithms are often the result of poor design or training of the machine. Societal biases are also introduced and rendered invisible when they are designed for profit-making alone. Then, society gets to bear the costs associated with them (Benkler, 2019).

Biased algorithms can result even when a very large amount of data is fed into the training process (Dahlmann & Dickow, 2019, p. 13).[12] If the machine is shown data extracted from social practices in which a discriminatory condition is embedded, this discrimination will be understood as legitimate and incorporated into its understanding of the model.[13]

Programming neural networks will also demand a deep understanding of the way they operate. Simple optimisation of the results will provide excellent solutions for the network at large, while potentially being unfair towards the least favoured individuals. The challenge is to teach algorithms to balance their decisions so that they achieve ethical goals even at the expense of optimal results. Balancing ethics and profit is not an equilibrium that all humans will find easy to agree upon.

The development of toolkits designed to mitigate discrimination and bias in machine learning (such as IBM's AI Fairness 360[14] or Aequitas[15]) point to the only solution that might ensure that optimisation is achieved at machine speed while fairness and explainability are also contemplated: tested algorithms controlling algorithms. A long process of refining these tools will nonetheless be necessary before they are able to achieve acceptable results.

3.6. *Explainability*

Sometimes referred to as explicability, intelligibility or even accountability, explainability is of paramount importance for most code developers as a necessary means to achieve the trust of the users (Newswise, 2013). The Defense Advanced Research Projects Agency (DARPA) has started a program on Explainable AI which strives to look inside the black boxes of algorithms and aims for full control of their decisions (Gunning, 2017).

Trustworthiness itself is one of the key concepts closely linked with explainability. Although not so prevalent in the codes included in this study, it is commonly found in the literature related to AI. The European Commission has recently published a report on ethical guidelines for a trustworthy AI (High-Level Expert Group on Artificial Intelligence, European Commission, 2019), stressing their importance.

Yet the very concept of explainability –and whether it should apply to AI systems or to algorithms– is vague. This is especially true if we want to establish a link between transparency and trust, in which public understanding of the underlying technology would be mandatory (Buiten, 2019, pp. 1–19). There are many technologies used in everyday life whose inner workings are not commonly understood, and yet they are regarded as safe based on the assumption that someone or something else is keeping an eye on them.

The EU's GDPR (General Data Protection Regulation) explores other avenues for enforcing explainability. It "provides an unambiguous 'right to explanation' with sweeping legal implications" whose "true power derives from its synergistic effects when combined with the algorithmic auditing and 'data protection by design' methodologies" (Casey, Farhangi, & Vogl, 2019). Instead of an individualised right to explanation, the GDPR charges DPAs (Data Protection Authorities) with auditing powers throughout all phases of algorithmic development and deployment (European Commission, 2018).

3.7. *LAWS in the codes of conduct*

Other than expressions of concern and generic calls for a ban on autonomous weapons –or corporate positions (as in Google (Pichay, 2018))–, only the IEEE and COMEST among the ethical codes not specifically devoted to LAWS develop the ethics of AI-based weapons systems further.

Regarding LAWS, Version #2 of the Ethically Aligned Design stresses the "additional ethical dimensions" involved in dealing with weapons systems as compared to autonomous systems that are not designed to cause harm. Human control of weapons systems and the accountability of both developers and operators on the one hand, and predictability and explainability on the other, are the differential factors between these two types of autonomous systems (IEEE, 2016).

For its part, COMEST engages in a discussion of the ethical and legal aspects of LAWS following a similar one on armed drones. It argues that fully autonomous weapons would violate IHL[16] as they "lack the main components required to ensure compliance with the principles of distinction and proportionality". It therefore recommends that human control be retained in all circumstances over these systems (COMEST, 2017).

[12] The large amount of data available forces "modern robotic systems (to) use sensor data fusion and information filtering". "If filters influence the information that reaches the operator or commander in such a way without being controllable by humans themselves, it is doubtful whether there is a significant level of control, and thus whether attributable decisions can be made in the field."

[13] Amazon's CV reader algorithm, for instance, mistook the proportion of males among the best performers in the company as a standard to be imitated. See https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[14] https://aif360.mybluemix.net/.

[15] http://www.datasciencepublicpolicy.org/projects/aequitas/.

[16] Others, like Ronald Arkin, believe that robots might prove more "humane" in war and adhere to IHL more consistently as they would not be subject to passions like fear (Arkin, 2018). A machine would, in his view, be better at sticking to the rules of engagement (RoE) than soldiers. Counterarguments to this include the technical inability of machines to positively identify combatants in complex environments.

3.8. LAWS vs. non-lethal weapons regulation

War is defined as a contest of wills.[17] As far as it achieves the desired effects in bending the adversary's resolve, anything can be considered a weapon. War in the 21st century is waged within the people. We become the battlefield as well as the weapon and the victim. War does not even target people's thoughts, but their feelings and sentiments. It does not have to affect reality, but just perceptions.

Algorithms are used ever more often to produce the same effects as traditional kinetic weapons. They come with a bonus of deniability, ubiquity, immediacy and, best of all, cleanliness. Responsibility is hard to assign and perceptions are bent in order to justify their effects.

Algorithms are also used in LAWS. So-called "killer robots" with varying degrees of autonomy are already in service in a number of armed forces around the globe. As these fall within the category of kinetic weapons, they are perceived to pose a clear and present danger and, therefore, tend to be scrutinised. Most of them, so far, are only targeting non-human objectives.

Although the United Nations CCW has been working on their definition and regulation since 2013, there is still today little consensus on what LAWS entail. The United States Department of Defense issued one of the earliest statements (US Department of Defense, 2012), but most nations and institutions have also introduced their own version (United Nations Institute for Disarmament Research, 2017), making it almost impossible to agree on a common understanding of the matter to be regulated.

AI systems may be associated with hardware and thus become the brains of robotic machines (LAWS). They may also remain in the domain of software (as chatbots) or be put into direct relationship with biological intelligence (brain–computer interface, BCI). All three forms of AI have already been present in our society for the last few years. Yet humans tend to be more aware of actions performed by the former, while minimising the influence of the other two.[18]

Beyond lethal autonomous robots, soft forms of AI present a real threat in cyberspace (Scharre, 2019). If an ever-increasing part of our lives takes place in cyberspace, a precautionary system equivalent to the physical one needs to be established for algorithms. A series of scenarios are conceivable regarding threats introduced by AI, be they in physical, digital or political security (Brundage et al., 2018).[19]

While these tools would not be capable of causing physical harm, they would have the potential to affect perceptions, becoming life-threatening for the user or allowing for a narrative to be developed around these false assumptions.

Other non-physical effects could also be made available to hostile or illegal actors through dual-use technologies. Anything from cybersecurity to information and disinformation campaigns will become commonplace the moment tools to deploy them become available as off-the-shelf products (Brundage et al., 2018).

4. Dual-use AI

Unlike with other technologies, the layout of artificial intelligence systems used as weapons is hard to distinguish from that designed for everyday beneficial uses. In practical terms, all the technologies associated with artificial intelligence have a dual utility, for civilian and military uses. "While they may serve legitimate societal objectives in some cases, they are also used to undermine human rights like freedom of expression and privacy" (Penney et al., 2018).

Therefore, decision-makers need to take special care when regulating the development of these technologies. The focus needs to be on controllability of the algorithms rather than on a less realistic intent to ensure the exclusiveness of their use by approved actors (European Commission, 2015b). There is an opportunity cost –both economic and social– for the slowdown in these systems' availability. Few would understand if exoskeletons that would allow a disabled person to regain some freedom of movement were not made available as soon as they were safe to use for their intended purpose. That same technology, however, could be used to enhance soldiers' capabilities on the battlefield.

Both this commercial rush and the military advantage such systems will entail when "weaponized" guarantee that they will be implemented and then used for hostile or illegal purposes –and used while still immature.[20] As with many other examples of technological developments throughout history, war provides the best excuse to implement newly introduced inventions that still need to be further tested for civilian use. The aircraft that were hastily brought into service in World War I serve as a good example.

Close competition between commercial companies in the marketing of products and their designers' lack of awareness of this potential dual-use prevent the implementation of security measures in the design of artificial intelligence systems as they are

[17] "War therefore is an act of violence to compel our opponent to fulfill our will." Carl Von Clausewitz, "On War".

[18] Some studies show that public opinion on the use of LAWS is prone to changing depending on the specific use of these weapons, and that acceptance of them might follow their actual coming into service. Contrary to what might be intuitive, there is no upfront opposition to the idea of autonomous weapons or, at the very least, this opposition cannot be considered a universal concept (Horowitz, 2016).

[19] Brundage et al. describe some scenarios in different domains. Some of them are actually happening today in some degree. In the field of digital security, they present threats arising from: automation of social engineering attacks, of service tasks in criminal cyber-offence, of vulnerability discovery and of hacking; human-like denial-of-service; automation; prioritisation of targets for cyberattacks using machine learning and exploitation of AI used in applications. A similar approach follows in the physical and political domains.

[20] The Committee of the UK's House of Lords on AI advises "that universities and research councils providing grants and funding to AI researchers must insist that applications for such money demonstrate an awareness of the implications of the research and how it might be misused, and include details of the steps that will be taken to prevent such misuse, before any funding is provided." (Select Committee on AI of the House of Lords, 2018).

increasingly regarded as a strategic resource (Fischer & Wenger, 2019). "The potential for well-meaning AI research to be used by others to cause harm is significant" (Select Committee on AI of the House of Lords, 2018).

By comparison, nuclear weapons and their associated technology for energy generation are perceived as dangerous and are therefore strictly regulated. The assets needed to develop them and the indiscrete infrastructure required for the assembly of nuclear weapons make them a matter of negotiation solely among states. The possibility of a direct application of AI technologies to weapons systems or, simply, to weakening the agency and autonomy of humans is far less obvious and therefore rarely taken into account (Boulanin, 2018). Both the development cost and the perceived cost associated with the misuse of digital weapons are far lower, if only because such misuse is not as obvious.

Unlike electronic warfare, which relies almost exclusively on equipment that is custom-made for defence purposes given the lack of a RoI (return on investment) on the civilian side, most AI designs are either first conceived for civilian use or find their way into the civilian market in their early stages (Edwards, Natalucci, Oberlin, & Tigkos, 2019). Most discouraging is the fact that these applications are sometimes developed for the entertainment industry and not for beneficial purposes. Examples include swarms of drones choreographed by an algorithm –as an equivalent to fireworks– with little or no regard to their potential for military action.

In-house solutions are seldom used by most militaries. In most nations, state-of-the-art research and development of military equipment has been transferred from the armed forces to universities and civilian corporations. Time and budget constraints lead to COTS (commercial off-the-shelf) products being incorporated into military equipment with little time or interest given to implementing an additional safety layer.

As a result, we are witnessing AI technologies that were originally designed for peaceful civilian uses being applied to weapons systems, to the great concern of many of the scientists who helped develop them in the first place. These commercial technologies are often either very easy to use or simply plug-and-play. This allows for the systems developed around them to be used by non-state actors such as terrorist groups with little or no training.

The concept of "dual-use research of concern" (DURC) applied by the Human Brain Project to neuroscience research has very good potential to be translated into AI R&D (Aicardi et al., 2018, pp. 1–21). The study provides several examples of technologies that could be used by the militaries.

The very existence of the technologies also amounts to a significant potential for misuse by non-state actors (Chertoff, 2018). Even if states were to refrain from using them, terrorist groups would not be deterred or constrained by ethical considerations. Therefore, technical solutions need to be found and implemented to deny their misuse or to make it so challenging that these organisations are effectively denied its usage.

Dual-use algorithms enable non-state actors to perform offensive activities formerly beyond their reach. Low-skilled individuals might be capable of targeting adversaries with technologies well above their expertise. Several scenarios are conceivable within this context. AI-controlled swarms, for example, could be employed in an autonomous mode even in electronically protected environments where remote communications are not possible. Terrorists –but also state actors– will find it less psychologically challenging to perpetrate these acts when physically and emotionally detached from action (Brundage et al., 2018).

This duality of use implies that technologies are, for the most part, ethically neutral. It is the way in which each actor utilises them that will have ethical or legal connotations. Therefore, ethics should target the users instead of the machines. The aim is not to design and build ethical algorithms but to prevent unintended uses of the systems in which they are embedded (European Commission, 2015a). Ethical concerns should be present in all phases of development, with a worst-case scenario mindset.

The democratisation of such a powerful tool/weapon needs to come with the ability to ensure the responsible use of these technologies. Flaws in design, a lack of safety mechanisms and purposeful illicit intent are possible causes of unintended consequences. If responsibility is to be shared among decision-makers, designers, manufacturers and operators, agency over the actions of the algorithms should be likewise shared. Kill switches or similar mechanisms should be available for any of them to terminate a process if need be.

In a man-made ecosystem, dual-use technologies designed and developed with civilian commercial use in mind become not only the tools, but also the domain on which militaries fight. As in cyberspace, this means that the battlefield is no longer a segregated scenario. There will still be a few dedicated networks, tactics, techniques and procedures but, for the most part, the militaries will not own the battlefield, the vectors (the tools used to deliver the payload) or even the data they use.

5. Guiding principles in light of the CCW's ethical code proposal

The deliberations of the United Nations CCW, beyond the inevitable dissent between powers, industries and political ideologies, have reached an agreement on a tentative set of guiding principles (Table 1). Their conclusions are very much in line with the principles contained in the "civilian" ethical codes. This forum is the sole venue in which states and non-state actors are officially discussing these topics.

Some of the guiding principles present in those "civilian" ethical codes are nonetheless difficult to reconcile with machines designed to cause harm:

5.1. Beneficence/relative beneficence

Usually the first and most common concept present in general-purpose ethical codes, beneficence is arguably not applicable to systems designed to cause harm. However, several states claim in the documents presented in Geneva that, although situations on the battlefield are violent, algorithms will be better at discriminating them and will therefore minimise the number of unwanted casualties,

Table 1
CCW's proposal of Guiding Principles. By the author, based on (CCW, 2018).

CCW Guiding Principles
1. IHL continues to apply fully to all weapons systems, including the potential development and use of lethal autonomous weapons systems.
2. Human responsibility for decisions on the use of weapons systems must be retained since accountability cannot be transferred to machines. This should be considered across the entire life cycle of the weapon system.
3. Accountability for developing, deploying and using any emerging weapons system in the framework of the CCW must be ensured in accordance with applicable international law, including through the operation of such systems within a responsible chain of human command and control.
4. In accordance with state obligations under international law, in studying the development, acquisition, or adoption of a new weapon, means or method of warfare, it must be determined whether its employment would, in some or all circumstances, be prohibited by international law.
5. When developing or acquiring new weapons systems based on emerging technologies in the area of LAWS, physical security, appropriate non-physical safeguards (including cyber-security against hacking or data spoofing), the risk of acquisition by terrorist groups and the risk of proliferation should be considered.
6. Risk assessments and mitigation measures should be part of the design, development, testing and deployment cycle of emerging technologies in any weapons systems.
7. Consideration should be given to the use of emerging technologies in the area of LAWS in upholding compliance with IHL and other applicable international legal obligations.
8. In crafting potential policy measures, emerging technologies in the area of LAWS should not be anthropomorphized.
9. Discussions and any potential policy measures taken within the context of the CCW should not hamper progress in or access to peaceful uses of intelligent autonomous technologies.
10. CCW (…) seeks to strike a balance between military necessity and humanitarian considerations.

thus complying with IHL (United States, 2018). It is therefore a relative beneficence that is presented in the case of LAWS.

The rationale behind this argument is that machines do not get tired, nor are they influenced by the environment. According to this line of thought, algorithms are much better at determining who or what is a legitimate target and at minimising collateral damage. The amount of data they can process being larger, they should be able to arrive at more precise and less biased conclusions. Time constraints would also be limited with a faster computing capacity.

A machine's lack of predictability is mitigated by the fact that the human thought process is no less obscure and unpredictable (McGrath & Gupta, 2018). The human decision-making process is less rational than is commonly assumed. This approach has nonetheless been repeatedly disputed by the ICRC (International Committee of the Red Cross, 2018). The distribution of workloads among several AI systems –of which one might obtain or correlate information, another would take the decision and yet a third could execute it– may further blur responsibility.

Most of the discussion currently taking place at the CCW is around definitions and minor details that will probably remain unresolved for some time (Dahlmann & Dickow, 2019). Some actors are asking for decisions to be frozen until a consensus is reached (Russian Federation, 2018). These problems do not, however, diminish the validity of and the consensus around the proposed principles.

Leading experts on this field advocate a less ambitious approach in which niche capabilities –like human targeting– are banned, instead of an unrealistic prohibition of intelligent weapons. While limiting the autonomy of algorithms in these cases would prevent machines from directly engaging humans, if the rest of the targeting process is left untouched in the hands of the AI system, the human operator might be left with too few options or be limited to a Go/No-Go decision. These weapons could escalate the conflict before human intervention is possible (Scharre, 2018, pp. 23–27).

The principle, therefore, seen from the CCW's perspective, could very well be transformed into "relative benefit" upon comparing human and machine results. This could also be applicable to other AI systems such as autonomous driving and allow for the deployment of fully autonomous vehicles on the basis that they present better casualties statistics. But, while fewer casualties might result from this line of action, decisions would be taken by agents other than humans. This thought is central to all discussions and closely related to human dignity.

5.2. Human dignity

Humans see themselves as especially worthy, and as deserving of preferential treatment. Regardless of whether that quality derives from a divine mandate or self-realisation, we all reject the idea of a human life being taken by someone or something not human. We see ourselves as superior to other beings.

By transferring agency to machines, humans become mere targets, data, objectives and objects. Deprived of power, humans lose their dignity and self-esteem. This, again, is applicable both to LAWS and to civilian uses of AI systems. It applies both to decisions on how to live (freedom) and to who should die (who/what takes the decision).

Allowing self-driving cars to decide whether to kill passengers or bystanders on their own (as in the "trolley problem" (Thomson, 1984)) is, in that regard, equivalent to letting an autonomous drone engage a target based on preprogrammed algorithms or, worse still, on self-trained ones. The very idea of unavoidable fate based on black and white decisions is repugnant to our thoughts. That is clearly shown in MIT's Moral Machine experiment,[21] in which we get to make decisions based on real scenarios that an autonomous vehicle finds. While that decision is usually straightforward, reversing the method and applying the resulting algorithm to different situations introduces a feeling of discomfort.

Should we consider it less of a breach of human dignity to allow algorithms to decide whom to kill than to have them filter the data we access and thence distort our perception of reality and our freedom to choose? If freedom and power are the basis of human dignity, then it is conceivable that the ethics on LAWS should extend also to non-lethal activities if they affect our dignity.

5.3. *Fairness*

While non-discrimination is most relevant in civilian oriented codes, respect for IHL regarding the principles of precaution, distinction and proportionality needs to be prioritised vis-à-vis the strict accomplishment of the mission objectives in military codes. China's Position Paper at the CCW's meeting in April 2018 acknowledged the difficulties of LAWS in sticking to those principles and in establishing accountability (China, 2018).

Attempts to use nuclear weapons treaties as a model[22] are likely to fail because LAWS are far more accessible to state and non-state actors than NBC weapons.[23] While nuclear weapons require resources that are usually only within the reach of a state, LAWS are relatively inexpensive tools. Their "democratisation" implies the need for a new model of governance.

The alternative example being proposed is, paradoxically, that of a relatively low-tech weapon: landmines, which, once seeded, remain out of further control by the owner (Gubrud, 2018). While the behaviour of mines is easy to predict, there remains the need to retain control over them, whether geographically or in terms of the possibility of deactivating them once they are no longer performing their intended role.

5.4. *Meaningful human control*

Control is the most prominent principle in all ethical codes. However, the conditions to be met in order to achieve "meaningful human control" are also a subject of debate. Even when explained in detail –as in the case of the International Committee for Robot Arms Control (ICRAC)– precise mechanisms to achieve such control remain elusive (Sharkey, 2018).[24] To avert this debate, others advise that the discussion should focus on the outcome rather than the process (Lewis, 2018, pp. 1–23). The presence of a human in or on the loop is therefore mandatory to ensure that the process does not preclude the outcome.

Finally, a joint proposal from Austria, Brazil and Chile moves that a legally-binding instrument ensures meaningful human control over, at least, critical functions in LAWS (Austria, Brazil, & Chile, 2018). Most often, these proposals advise that humans should be the ones "pulling the trigger", while being far more permissive with the decision-making process. In my view, this places the responsibility on a usually ill-informed shooter, who will have little choice but to trust the algorithms that have selected the target.

Table 2 offers a comparison of the different interpretations of "meaningful human control". They all revolve around the idea of the ability of the operators to understand the technology and to intervene, in the event that the outcome is not aligned with the intended one. This understanding implies the need for the operator to be knowledgeable of the context and ready to incorporate the information provided by the algorithms. Meaningful human control, therefore, advocates the use of AI in a supporting role, automating processes but keeping the human in the loop at each step.

In the classification proposed in Table 3, Parasuraman, Sheridan, and Wickens (2000) present the decreasing levels of automation of decision and action. As discussed above, decisions taken by algorithms might prevent the operator from even considering some options, thus conditioning his agency anywhere from as low as level 3.

Noel Sharkey (Sharkey, 2018) reduces the list to 5 levels (Table 4), deeming level 3 and beyond as unacceptable for those same reasons. Target selection by the algorithm, even when the decision whether to strike or not is taken by the operator, greatly conditions his or her understanding of the whole picture and ability to make an informed decision. Level 4, for example, limits human intervention for approval to a specified period of time. These last two options would be open only in limited scenarios in which all targets are legitimate and the only goal is the prioritisation of the engagement.

Such scenarios may include the likes of outer space operations, deep sea submarine warfare or the protection of critical

[21] http://moralmachine.mit.edu/.

[22] Sharikov, nonetheless, proposes the Anti-Ballistic Missiles Treaty (ABM) as a possible model (Sharikov, 2019).

[23] NBC- nuclear, biological or chemical weapons. Notwithstanding the differences with nuclear weapons, there are those who believe that the deterrent effect of non-recallable weapons might be comparable with MAD (Mutual Assured Destruction), which helped prevent the use of atomic weapons during the Cold War. Compounded with the obvious tactical benefits of these technologies, LAWS might become unavoidable (Straub, 2016).

[24] According to ICRAC: Necessary conditions for meaningful human control of weapons. A commander or operator should: 1. have full contextual and situational awareness of the target area at the time of initiating a specific attack; 2. be able to perceive and react to any change or unanticipated situations that may have arisen since planning the attack, such as changes in the legitimacy of the targets; 3. have active cognitive participation in the attack; 4. have sufficient time for deliberation on the nature of targets, their significance in terms of the necessity and appropriateness of an attack, and the likely incidental and possible accidental effects of the attack; and 5. have a means for the rapid suspension or abortion of the attack (Asaro, 2019; Sharkey, 2018).

Table 2

A comparison of different criteria regarding the significance of "meaningful human control". By the author, based on (Ekelhof, 2018). CNAS (Centre for a New American Security), Article 36 Geneva Conv. (Article 36 of the 1977 Additional Protocol to the 1949 Geneva Conventions), ICRAC (International Committee for Robot Arms Control) and ICRC (International Committee of the Red Cross).

	CNAS	Article 36 Geneva Conv.	ICRAC	ICRC
Human participation	Informed conscious decisions	Timely human judgement and action	Cognitive participation. Perceive and react	Human intervention in all stages
Information needed	Sufficient information on the weapon, the target and the context	Accurate information on technology, objective and context	Nature of target and collateral damage. Full contextual and situational awareness	Information about the weapons system and the context
Design of weapon	Weapon tested. Human trained	Predictable, reliable and transparent technology.	Suspension/abortion of attack.	Predictability and reliability.
Legal requisites	Enough information to ensure lawfulness	Accountability to a certain standard.	Necessity and appropriateness of attack. Compliance with IHL	Accountability and compliance with IHL

Table 3
A model for types and levels of human interaction with automation. Compiled by the author, based on (Parasuraman et al., 2000).

1	Human takes all decisions and actions
2	Computer offers alternatives, human decides and executes
3	Computer narrows alternatives, human decides on available ones
4	Computer offers one alternative, human approves and executes
5	Computer offers one alternative, human approves, computer executes
6	Computer decides and executes, human has limited time to veto
7	Computer executes and informs human
8	Computer executes and informs human on demand (pull)
9	Computer executes and informs human if it decides to (push)
10	Computer decides, executes and ignores human

Table 4
Levels of human control and how they impact on human decision-making. By the author, based on (Sharkey, 2018).

	IDENTIFICATION	SELECTION	APPROVAL	ATTACK
1	Human	Human	Human	LAWS
2	LAWS	Human	Human	LAWS
3	LAWS	LAWS	Human	LAWS
4	LAWS	LAWS	Human (time)	LAWS
5	LAWS	LAWS	LAWS	LAWS

infrastructure within a given perimeter. The scenarios do not necessarily need to be geographical; the speed of the target, its chemical composition and other factors may very well exclude that human lives are at risk. In both types of scenarios, geographical and otherwise, level five autonomy might be permissible and even advisable as further delays introduced by human action might endanger humans at the defending end.

It is human autonomy that is relevant in the ethical debate, however. Benefit, human dignity and justice are associated with control over machines. Agency is a rivalrous good that can only be enjoyed by one agent, so that the transfer of decisions to algorithms deprives humans of their ability to exercise such agency. The key lies in the definition of the boundaries between "assistance with the decision" and "conditioning of the narrative".

Privacy and explainability are also closely related ideas. AI should allow us to understand how machines work and the way in which it makes decisions while safeguarding human privacy. Transparency should apply to algorithms while humans retain ownership over their data; that is, turning machines into transparent enough tools while avoiding making users transparent.

Therefore, codes of conduct are about understanding algorithms (explainability) and being able to control them. Humans need to retain both power and responsibility (agency) (Boulanin, 2018). The aim of AI should be for it to be beneficial for individual humans (human dignity and privacy), while providing the tools for governance for societies (fairness), but not at the expense of the individuals.

6. Discussion and conclusion

AI provides undoubted benefits to humanity, the development of which should not be jeopardised by the need to introduce security measures that mitigate the risks associated with its deployment.

Among these threats, the dual use of these technologies takes a prominent place. The possibility of these systems being used by states or non-state actors for purposes contrary to national law or IHL must be taken into account in their design, development and use, but should not prevent them.

Autonomy, not lethality, is the key factor affecting human dignity. Any degree of unsupervised autonomy for machines happens at the expense of human autonomy, freedom and power.

Algorithms that condition freedom, albeit without exerting kinetic action on the target, should be considered instruments of war and treated as weapons. Hence, ethical standards applicable to LAWS should also be taken into consideration when designing, developing or utilising other AI systems that interact with human volition and liberty. Any type of coercion, be it physical or logical, should be interpreted as a hostile act, regardless of whether it involves lethality.

Therefore, principles initially drafted for hard forms of AI like LAWS are equally valid for soft AI if the latter is used for determining or influencing decisions in a way that may deprive humans of their agency. There should be no question about the need to incorporate "military grade" ethical standards into civilian technologies that are susceptible to being used as part of a weapons system.

Retaining control over the decisions is therefore the key principle, as it implies that the ethical debate remains with the user, the human designer and/or operator. Explainability is also mandatory as, in order to exercise control, we need a full understanding of the inner workings of algorithms.

Even if AI is very useful in a supporting role and may benefit humans, responsibility needs to remain with us. Power and responsibility need to stay together, and we cannot afford to cede power simply avoid taking responsibility (Belgium et al., 2019; Docherty & Human Rights Watch, 2019).

Appendix

SUMMARY OF KEY PRINCIPLES IN AI ETHICAL CODES

	BENEFICENCE/ HUMAN DIGNITY	PRIVACY	HUMAN AUTONOMY	FAIRNESS	EXPLAINABILITY
AI4PEOPLE	Beneficence: Promoting Well-Being, Preserving Dignity, and Sustaining the Planet	Non-maleficence/Privacy/ Security	Autonomy and human agency	Justice: Promoting Prosperity and Preserving Solidarity	Explicability/Intelligibility/ Accountability
EAD2	Prioritizing Well-being		Human Rights		Transparency/ Accountability
ASILOMAR	Value Alignment/ Human Values	Safety/Personal Privacy/ Liberty and Privacy/Non-subversion	Human Control	Shared Benefit/ Shared Prosperity	Failure Transparency/ Judicial Transparency/ Responsibility
MONTREAL	Well being	Protection of privacy and intimacy	Human autonomy/ Responsibility	Solidarity/Equity/ Diversity inclusion	Democratic participation
COMISION UE	Respecting the refusal of care by a robot	Protecting humans from harm/ liberty/against privacy breaches/Managing personal data processed by robots	Protecting humanity against the risk of manipulation by robots	Avoiding the dissolution of social ties/Equal access to progress in robotics	
UK LORDS		Intelligibility/Explainability/ Transparency/ Anonymisation./Liability/ criminal misuse		Education/social and political cohesion/ inequality	
PARTNERSHIP	Benefit and empower as many people as possible	Open research of legal consequences/Privacy and security of individuals/AI research and technology is robust, reliable, trustworthy, and operates within secure constraints	Human Rights	Social responsibility/ Culture of cooperation, trust, and openness	Educate and listen to public/Accountability/ Understandable and interpretable
EPSRC	Robots are manufactured artefacts: the illusion of emotions and intent should not be used to exploit vulnerable users.	Compliance/Privacy/Safety and security			Responsibility
COMEST	Beneficence/Human Dignity	Privacy/Do not harm	Human autonomy	Responsibility (liability, transparency, accountability)/ Cultural diversity/ Justice (inequality)	
UE	Dignity/Cumulative knowledge	Identity/Privacy and data protection/Surveillance and datafication/Democracy and trust (profiling, bots, fake news, freedom of expression)	Human Autonomy	Fairness and equity	Responsibility, accountability and transparency
FAT/ML		Responsibility		Fairness	Explainability/Accuracy/ Auditability
GOOGLE	Socially beneficial	Safety/Privacy		Avoid creating unfair byas/Be accountable to people	
BARCELONA		Prudence/Reliability	Responsibility	Human role	Accountability
UNI Global Union	Serve people and the planet	Transparency/Traceability/ Control	Fundamental Freedoms and Rights/Ban of Robots's responsibility	Equity/Unbyas	Global Governance Mechanisms
ACM	Society and human well- being/Intellectual property/general good	Avoid harm/Privacy		Fairness/Non discrimination	Honesty/Trustworthyness
IBM	enhance and extend human capability, expertise and potential	Safety/Security	Human control		Transparency
DEEPMIND	Social benefit	Transparency			

(*continued on next page*)

(*continued*)

	BENEFICENCE/ HUMAN DIGNITY	PRIVACY	HUMAN AUTONOMY	FAIRNESS	EXPLAINABILITY
Microsoft		Reliability/Safety/Security		Fairness/ Inclusiveness	Transparency/ Accountability
EURON	Human dignity/ Benefit/ Biosustainability	No harm/Informed consent/ Privacy/confidentiality	Human rights/ Autonomy and individual responsibility	Justice, equality, equity/Cultural diversity and pluralism/Non discrimination, non stigmatization	Accountability/ Transparency/Act + explain
BAAI Principles	Promote the sustainable development of nature and society	Be responsible, control risks.	Conform to human values as well as the overall interests of humankind.	Human privacy, dignity, freedom, autonomy, and rights should be sufficiently respected.	Be Ethical. Trustworthyness, fairness, reducing discrimination and biases, transparency, explainability, predictability, traceability, auditability and accountability.
DARPA		Trust/Explainability/ Traceability			
CCW	(Recomendations) Human responsibility and accountability/ Human control	Do not hamper progress or peaceful uses	Safety and security of systems, be it physical or logical/Risk assesment and mitigation measures/Non	antropophormization	
OECD	Inclusive growth, sustainable development and well-being	Robustness, security and safety	Human-centred values and fairness	Accountability	Transparency and explainability

References

ACM. (2018). *Code of ethics and professional conduct* (pp. 1–42). Association for Computing Machinery. Retrieved from https://www.acm.org/about-acm/acm-code-of-ethics-and-professional-conduct#CONTENTS.

Aicardi, Christine, Bitsch, Lise, Badum, Nicklas B., Datta, Saheli, Evers, Kathinka, Farisco, Michele, et al. (2018). *Opinion on "responsible dual use". Human Brain Project*. Danish Board of Technology Foundation. Retrieved from https://sos-ch-dk-2.exo.io/public-website-production/filer_public/f8/f0/f8f09276-d370-4758-ad03-679fa1c57e95/hbp-ethics-society-2018-opinion-on-dual-use.pdf. In press.

Arkin, R. (2018). Lethal autonomous systems and the plight of the non-combatant. In *The political economy of robots* (pp. 317–326). Cham: Palgrave MacMillan. https://doi.org/10.1007/978-3-319-51466-6_15.

Asaro, P. M. (2019). *ICRAC statement at the March 2019 CCW GGE*. CCW. Retrieved from https://www.icrac.net/icrac-statement-at-the-march-2019-ccw-gge/.

Austria, Brazil, & Chile. (2018). Proposal for a mandate to negotiate a legally-binding instrument that addresses the legal, humanitarian and ethical concerns posed by emerging technologies in the area of lethal autonomous weapons systems (LAWS). In *2018 meeting of the high contracting parties to the CCW*. Retrieved from http://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2018/gge/documents/29August_Proposal_Mandate_ABC.pdf.

Barcelona Bar Association. (2019). *Barcelona declaration*. Retrieved from https://www.iiia.csic.es/barcelonadeclaration/.

Baura, G. D. (2006). *Engineering ethics: An industrial perspective*. Elsevier Academic Press.

Beijing Academy of AI. (2019). *Beijing AI principles* - 新闻. Retrieved June 1, 2019, from https://www.baai.ac.cn/blog/beijing-ai-principles.

Belgium, Ireland, & Luxembourg. (2019). Food for thought paper. In *4th session of the GGE LAWS* (pp. 1–7).

Benkler, Y. (2019). Don't let industry write the rules for AI. *Nature, 569*, 161. Retrieved from https://www.nature.com/articles/d41586-019-01413-1.

Boddington, P. (2017). *Towards a code of ethics in artificial intelligence*. Retrieved from https://futureoflife.org/2017/07/31/towards-a-code-of-ethics-in-artificial-intelligence/.

Boden, M., Bryson, J., Calwell, D., Dauthenhahn, K., Edwards, L., Kember, S., et al. (2010). *Principles of robotics*. Retrieved from https://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics/.

Bolivarian Republic of Venezuela. (2018). General principles on lethal autonomous weapons systems. In *Group of governmental experts of the high contracting parties to the CCW* (pp. 1–2).

Boulanin, V. (2018). *The impact of artificial intelligence on strategic stability and nuclear risk*. SIPRI. https://doi.org/10.2991/icsshe-18.2018.203.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., et al. (2018). *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation (February 2018)*. https://doi.org/10.1002/adma.201405087.

Buiten, M. C. (2019). *Towards intelligent regulation of artificial intelligence*. European Journal of Risk Regulation. https://doi.org/10.1017/err.2019.8. October 2018.

Campaign to Stop Killer Robots. (2018). *Campaign to Stop killer robots*. Retrieved from https://www.stopkillerrobots.org/.

Casey, B., Farhangi, A., & Vogl, R. (2019). Rethinking explainable machines. *Berkeley Technology Law Journal, 34*, 1–50. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3143325.

CCW. (2018). *Report of the 2018 group of governmental experts on lethal autonomous weapons systems*.

Chertoff, A. P. (2018). *Perils of lethal autonomous weapons systems proliferation: Preventing non-state acquisition*. Geneva Centre for Security Policy (2). Retrieved from https://www.gcsp.ch/News-Knowledge/Publications/Perils-of-Lethal-Autonomous-Weapons-Systems-Proliferation-Preventing-Non-State-Acquisition.

China. (2018). Position paper. In *2018 meeting of the high contracting parties to the CCW* (pp. 1–22).

China-UK Research Centre for AI Ethics and Governance. (2018). Linking Artificial Intelligence Principles (LAIP) (n.d.)Retrieved June 11, 2019, from http://www.linking-ai-principles.org/principles.

COMEST. (2017). *Report of COMEST on robotics ethics*. Paris. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000253952.

Craglia, M. (2018). *Artificial intelligence: A European perspective - European Commission*. Luxembourg. Retrieved from https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/artificial-intelligence-european-perspective.

Dahlmann, A., & Dickow, M. (2019). *Preventive regulation of autonomous weapon systems: Need for action by Germany at various levels*. German Institute for International and Security Affairs (3). Retrieved from https://www.ssoar.info/ssoar/bitstream/handle/document/62252/ssoar-2019-dahlmann_et_al-Preventive_regulation_of_autonomous_weapon.pdf?sequence=1&isAllowed=y&lnkname=ssoar-2019-dahlmann_et_al-Preventive_regulation_of_autonomous_weapon.pdf.

DeepMind. (2017). *DeepMind ethics and society principles*. Retrieved from https://deepmind.com/applied/deepmind-ethics-society/principles/.

Defense Innovation Board. (2019a). *AI principles: Recommendations on the ethical use of artificial intelligence by the*. Department of Defense.

Defense Innovation Board. (2019b). *AI principles: Recommendations on the ethical use of artificial intelligence by the department of Defense*. Defense Innovation Board Supporting Document.

Diakopoulos, N., Friedler, S., Arenas, M., Barocas, S., Hay, M., Howe, B., et al. (2018). *Principles for accountable Algorithms and a social impact statement for algorithms*. http://www.fatml.org/resources/principles-for-accountable-algorithms. Retrieved January 12, 2019, from.

Docherty, B., & Human Rights Watch. (2019). Agenda item 5(a) regarding challenges posed to international humanitarian law. In *Statement to convention on conventional weapons (CCW) group of governmental experts on lethal autonomous weapons systems agenda* (p. 2).

Edwards, J., Natalucci, M., Oberlin, J., & Tigkos, K. (2019). *C4ISR and network centric warfare: Current trends and projected developments content*. Jane's Defence Industry & Markets Intelligence Centre.

Ekelhof, Merel A. C. (2018). Lifting the fog of targeting: "Autonomous Weapons" and human control through the lens of military targeting. *Naval War College Review, 71*(3), 6.

Ekelhof, M. A. C. (2017). Complications of a common language: Why it is so hard to talk about autonomous weapons. *Journal of Conflict and Security Law, 22*(2), 311–331. https://doi.org/10.1093/jcsl/krw029.

European Commission. (2015a). *Explanatory note on potential misuse of research*.

European Commission. (2015b). Research with an exclusive focus on civil applications only. *Pharmaceuticals Policy and Law, 6*.

European Commission. (2018). *General data protection regulation (GDPR)*. Retrieved June 4, 2019, from https://gdpr-info.eu/.

European Parliament Committee on Legal Affairs. (2016). *Civil law rules on robotics*. Retrieved from http://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU(2016)571379_EN.pdf.

Fischer, S.-C., & Wenger, A. (2019). *A politically neutral hub for basic AI research. CSS ETH Zurich Policy Perspectives, 7(March)*.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., et al. (2018). AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5.

Future of Life Institute. (2017). Asilomar principles - future of life Institute. In *2017 asilomar conference*. Retrieved from https://futureoflife.org/ai-principles/.

Gubrud, M. A. (2018). The Ottawa definition of landmines as a start to defining LAWS. In *Convention on conventional weapons group of governmental experts meeting on lethal autonomous weapons systems, (April)*.

Gunkel, D. J. (2018). The other question: Can and should robots have rights? *Ethics and Information Technology, 20*(2), 87–99. https://doi.org/10.1007/s10676-017-9442-4.

Gunning, D. (2017). *Explainable artificial intelligence (XAI). The need for explainable AI*. https://doi.org/10.1111/fct.12208 (November).

High-Level Expert Group on Artificial Intelligence, European Commission. (2019). *Ethics guidelines for trustworthy AI*. Retrieved from https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=58477.

Horowitz, M. C. (2016). Public opinion and the politics of the killer robots debate. *Research & Politics, 3*(1). https://doi.org/10.1177/2053168015627183, 205316801562718.

IBM. (2018). *IBM partnerworld code of conduct*. Retrieved from https://www.ibm.com/partnerworld/program/code-of-conduct.

IEEE. (2016). *Ethically aligned design V2*. https://doi.org/10.1109/MCS.2018.2810458.

International Committee of the Red Cross. (2018). Ethics and autonomous weapon systems: An ethical basis for human control?. In *Group of governmental experts of the high contracting parties to the CCW* (pp. 1–22).

Lewis, L. (2018). *Redefining human control. Lessons from the battlefield for autonomous weapons*. Center for Autonomy and AI (Retrieved from cna.org/CAAI).

Leys, N. (2018). Autonomous weapon systems and international crises. *Strategic Studies Quarterly, 12*(1), 48–73. https://doi.org/10.2307/26333877.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes, 151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005.

McGrath, J., & Gupta, A. (2018). Writing a moral code: Algorithms for ethical reasoning by humans and machines. *Religions, 9*(8), 240. https://doi.org/10.3390/rel9080240.

Microsoft. (2019). *AI principles*. Retrieved from https://www.microsoft.com/en-us/ai/our-approach-to-ai.

Muehlhauser, L., & Helm, L. (2012). Intelligence explosion and machine ethics. *Singularity Hypothesis: A Scientific and Philosophical Assessment, 6*, 1–28. https://doi.org/10.1007/978-3-642-32560-1_6.

Newswise. (2013). *New survey shows widespread opposition to 'killer robots,' support for new ban campaign*. Retrieved April 27, 2019, from https://www.newswise.com/articles/new-survey-shows-widespread-opposition-to-killer-robots-support-for-new-ban-campaign.

OECD. (2018). AI: Intelligent machines, smart policies. In *OECD digital economy papers*. Retrieved from https://artsandculture.google.com/.

OECD. (2019a). *OECD principles on artificial intelligence*. Retrieved June 4, 2019, from https://www.oecd.org/going-digital/ai/principles/.

OECD. (2019b). *Recommendation of the council on artificial intelligence*. Environment (April).

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 30*(3), 286–297. https://doi.org/10.1109/3468.844354.

Penney, J., McKune, S., Gill, L., & Deibert, R. J. (2018). Advancing human-rights-by-design in the dual-use technology industry. *Colombia Journal of International Affairs*, (April). Retrieved from https://jia.sipa.columbia.edu/advancing-human-rights-design-dual-use-technology-industry.

Pichay, S. (2018). *Our principles – Google AI*. Retrieved April 27, 2019, from https://ai.google/principles/.

Potoglou, D., Patil, S., Gijon, C., Palacios, J., & Feijoo, C. (2013). The value of personal information online: Results from three stated preference discrete choice experiments in the UK. ORCA. In *21st European conference for information systems. Utrecht, The Netherlands*. Retrieved from http://orca.cf.ac.uk/51292/.

Rothenberger, L., Fabian, B., & Arunov, E. (2019). Relevance of ethical guidelines for artificial intelligence – a survey and evaluation. In *Proc. 27th European conference on information systems, stockholm & Uppsala, Sweden, June 8-14, 2019*.

Russian Federation. (2018). Russia's approaches to the elaboration of a working definition and basic functions of lethal autonomous weapons systems in the context of the purposes and objectives of the convention. In *Group of governmental experts of the high contracting parties to the CCW* (pp. 1–3).

Scharre, P. (2018). *A million mistakes a second. Foreign Policy, Fall*.

Scharre, P. (2019). *Killer apps: The real dangers of an AI arms race. Foreign Affairs, (June), 135–145*. Retrieved from https://www.foreignaffairs.com/articles/2019-04-16/killer-apps.

Select Committee on AI of the House of Lords. (2018). *AI in the UK: Ready, willing and able?*. Retrieved from https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10001.htm.

Sharikov, P. (2019). *Artificial intelligence, cyberattacks and nuclear weapons: A dangerous combination*. EastWest Institute. Retrieved from https://www.eastwest.ngo/idea/artificial-intelligence-cyber-attacks-and-nuclear-weapons-dangerous-combination.

Sharkey, N. (2018). *Guidelines for the human control of weapons systems*. Retrieved from http://bit.ly/1h6X6jB.

Straub, J. (2016). Consideration of the use of autonomous, non-recallable unmanned vehicles and programs as a deterrent or threat by state actors and others. *Technology in Society, 44*, 39–47. https://doi.org/10.1016/J.TECHSOC.2015.12.003.

Tavani, H. T. (2018). Can social robots qualify for moral consideration? Reframing the question about robot rights. *Information, 9*(4), 73. https://doi.org/10.3390/info9040073.

The Partnership on AI. (2018). *Tenets*. Retrieved January 3, 2019, from https://www.partnershiponai.org/tenets/.

Thomson, J. J. (1984). *The trolley problem* (Vol. 94). Yale Law Journal. Retrieved from https://heinonline.org/HOL/Page?handle=hein.journals/ylr94&id=1415&div=&collection=.

Williams, Andrew P., & Scharre, Paul D. (2015). Autonomous systems issues for defence policymakers. *NATO ACT*. https://apps.dtic.mil/dtic/tr/fulltext/u2/1010077.pdf.

Trinity International University (n.d.). The Center for Bioethics & Human Dignity. Retrieved May 19, 2019, from https://cbhd.org/category/issues/human-dignity.

UNI Global Union. (2018). *10 principles for ethical AI*. Retrieved February 1, 2019, from http://www.thefutureworldofwork.org/opinions/10-principles-for-ethical-ai/.

United Nations Institute for Disarmament Research. (2017). The weaponization of increasingly autonomous technologies: Concerns, characteristics and definitional approaches, a primer. *UNIDIR Resources, 6*(6). Retrieved from www.unidir.org.

University of Montreal. (2018). *Montreal declaration for the responsible development of AI*. Retrieved January 3, 2019, from https://www.montrealdeclaration-responsibleai.com/the-declaration.

Upshaw Downs, J. C., & Swienton, A. R. (2012). *Ethics in forensic science*. Academic.

US Department of Defense. (2012). *Autonomy in weapon systems*. Retrieved from https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf.