# THE ETHICS OF ARTIFICIAL INTELLIGENCE AND THE MULTILATERAL PUSH FOR A TREATY

**Master Thesis in International Affairs and Diplomacy**

**Course 2021-2022**

**Pablo González Peralta**

Thesis supervisor: Just Castillo Iglesias

March 2022

**Word count: 17,640**

# ABSTRACT

The technologies embedding Artificial Intelligence (AI) algorithms have had, in the recent years, an exponential development both in terms of capabilities and reach. AI has already contributed to spectacular advances in many technological and scientific areas. In parallel, AI applications can also be used to manipulate citizens, amplify cyberattacks, abuse biometric data, constrain, or cancel, freedoms and control lethal autonomous weapons.

As the world has known better the risks associated to AI both at societal and at military levels, the need to settle ethical principles that pose limits to the development, management and usage of AI has become urgent. This article examines the journey that different international organizations have followed in recommending AI normative frameworks based on ethical principles. The proposed ethical principles are described and discussed, as well as the need to move from recommendations to laws, as illustrated by the EU AI Act.

AI-based systems create new threats to the global security. Balances of power, new security dilemmas, autonomous weaponry and increased uncertainty are some of the elements discussed in this article. The possibilities for AI to launch and lead 'hyper' and 'cyber' wars pose tremendous dangers to humanity. There is a growing consensus that those risks should be contained. International law should prevail, ethical principles respected, and humans need to remain accountable and in control.

The international community is ready to work in a treaty to regulate the military applications of AI. This document argues that a treaty is needed, needed now, and explores some of the proposed foundational elements that such treaty should consider.

**Keywords:**

Artificial Intelligence, AI, Ethics, LAWS, Autonomous Weapons, International Humanitarian Law, International Law, International Security, International Treaty, Cyberspace, Hyperwar, Multilateralism, OECD, UNESCO

# TABLE OF CONTENTS

# DEFINITIONS

*Source: "Recommendation of the Council on Artificial Intelligence", OECD/LEGAL/0449, OECD, 2019*

***AI system**: An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.*

***AI system lifecycle**: AI system lifecycle phases involve: i) 'design, data and models'; which is a context-dependent sequence encompassing planning and design, data collection and processing, as well as model building; ii) 'verification and validation'; iii) 'deployment'; and iv) 'operation and monitoring'. These phases often take place in an iterative manner and are not necessarily sequential. The decision to retire an AI system from operation may occur at any point during the operation and monitoring phase.*

***AI actors**: AI actors are those who play an active role in the AI system lifecycle, including organisations and individuals that deploy or operate AI*

# INTRODUCTION

> *The "Artificial Intelligence" term was coined in 1955, by John McCarthy, Marvin L. Minsky, Nathaniel Rochester and Claude E. Shannon. The "study of artificial intelligence" was planned "to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it".*
>
> *As the field developed and diversified in the decades to come, the number of meanings of "AI" increased and there is no universally agreed upon definition today. Various definitions of AI are related to different disciplinary approaches such as computer science, electrical engineering, robotics, psychology, or philosophy. (UNESCO 2019)*

As seen above, the theoretical framework of AI dates from mid of the last century but the current explosion of AI developments and applications are the consequence of three technological mega-trends: big data, machine learning, and supercomputing. The advances in computer science, in robotics, the availability of new forms and huge amounts of data, the sensors and cameras becoming cheap and ubiquitous and the integrated and interconnected world where we live in, are all factors that have contributed to this revolution.

As for any new disruptive technology, the application of AI algorithms will impact both the civilian and the military worlds as we know them today. AI has already contributed to spectacular advances in medical science, autonomous vehicles, industrial robots, and climate change analysis among other fields. If correctly handled and shared, these advances can contribute to create a better world, more sustainable, more equitable with higher degrees of wellbeing.

At the same time, this technology brings new moral and ethical challenges. AI can heavily erode democracies, strengthen and sustain authoritarian regimes, and become a security destabilizing force. Current AI systems based on big data can twist established citizens' rights and freedoms, can help manipulate people along obscure political purposes, can track people, and discriminate through social scoring. In military applications, AI systems can autonomously detect targets to eliminate or command swarms of drones without human direct control.

The Alan Turing Institute in 2021 warns of the risks to human rights and freedoms of AI in areas such the judicial system, privacy, freedom of expression, equality, and social and economic rights. Surveillance on the work environment or AI powered decision on hiring or pay will affect economic opportunities of heavily biased profiles (ATI 2021). Our societies may not be prepared for the changes that will follow and we will need to quickly adapt to this new reality and become knowledgeable on how AI works, and the risks associated.

The countries that lead the AI race will enjoy huge economic competitive advantages over the rest if nothing is done to share its benefits. Economic power translates into military power. In Buchanan (2020), machine learning is defined as a system using "computing power to execute algorithms that learn from data". Each component of this "AI Triad", algorithms, data, and computing power, should be moved upwards in the policy priority of any country.

On the security side, AI can potentially destabilize the current balance of power as the US and China have well understood. How to guarantee principles like accountability,

transparency and human oversight and ultimate control, are common themes of concern in the conflicts to come where AI can play an important role.

AI policy making, and governance should also extend beyond the state. On one hand, large companies dominate most of the AI development and biggest innovations but the barrier of entry for this technology is relatively low and multitude of start-ups focus on niche applications that could be potentially harmful and less controlled. An effective governance would probably require multi-stakeholder approaches involving the civil society and the private sector.

In addition, the asymmetrical power that countries that dominate this technology may have over those that do not have enough capacity to develop it cannot be counterbalanced without the deep involvement of multilateral international institutions like the United Nations (UN).

As warned by Pauwels (2018), powerful nations but also large technology platforms could compete for our collective data for supremacy with the risk of enabling what begins to be called "cyber-colonization" whereby AI powered states could potentially control other countries' populations and ecosystems.

While the principles of democracy and human rights have different modulations across the globe, some recent efforts carried out by the UN, and the UNESCO in particular, to level up the ethics of AI are worth to mention and will be the subject of this study.

Through a descriptive approach, this thesis tries to understand the efforts made by different scholars, international organizations, and lead scientists in providing a regulatory framework and policy mechanisms to address the growing innovation that AI is bringing at exponential speed. The ethical and human challenges, risks and security threats faced when using AI in both civilian and military applications are described. The usage of AI will shamble the current balance of power, create new security dilemmas and introduce new considerations to factor in international law. The thesis finally explores the need to work on an international treaty that could limit or forbid the most harmful effects of AI in armed conflicts while building trust between the nations.

The last six months have been rich in the publication of new ethical norms and recommendations. The sources used to write this thesis include publications of international organisms like the OECD, EU, NATO and UNESCO, and the work of scholars, journalists and think tanks specialised on the topic.

The reminder of this thesis is structured as follows. The first chapter explores the normative landscape around AI ethical principles, discusses the importance to develop an ethical AI and argues that the current recommendations and guidelines need to evolve into national laws monitored by supervisory bodies. The second chapter analyses the threats to the global security that AI is creating or amplifying. It highlights how this new technology is changing the structure of the armed conflict sending shock waves to the current security order. Finally, the third chapter discusses the role of multilateralism in addressing the threats identified, argues the need for an AI treaty and discusses some of the proposed elements that it should incorporate.

# CHAPTER 1: The Journey towards a Global AI Ethics Norm

## 1.1. The need for AI Norms, Values and Principles

The debate around the need to set ethical AI principles started as soon as the technology was mature enough to unveil the risks associated with using it. Those risks, among others, included the accountability associated with systems that would make decisions on their own without human intervention, the risks of making biased decisions product of non-transparent development and training cycles, or the risk of using AI system with malicious objectives.

In 2015, the Future of Life Institute, founded by MIT professor Max Tegmark and funded among others by Elon Musk, reunited dozens of researchers in Puerto Rico with the goal to debate on how to keep AI beneficial for humanity. 37 working groups across the world received funds to research in that direction. As a follow up of Puerto Rico, in 2017, the Future of Life Institute organized in Asilomar, California, the Beneficial Artificial Intelligence 2017 conference reuniting leading AI researchers. AI leaders came together to discuss opportunities and challenges related to the future of AI and steps that could be taken to ensure that the technology is beneficial.

The Asilomar conference agreed on 23 ethical principles for AI that were grouped into three categories: Research Issues, Ethics and Values, and Longer-Term Issues. While those principles were just succinctly defined, they included already a plea to avoid an AI arms race, the need to respect human dignity, rights and freedoms, and a demand on making sure that humans choose how and whether to delegate decisions to AI (FoL 2017).

Those principles were endorsed so far by 1,800 AI researchers and more than 3,900 scientific personalities including Stephen Hawking, Elon Musk, and Jaan Tallinn.

In 2019, the Beijing Academy of Artificial Intelligence (BAAI) led the development of 15 AI principles, endorsed by the major universities and national research institutions in China (IRC 2019). All the major private companies worldwide that research and develop AI systems have crafted and published their own ethical principles and they claim they abide to them. In a work from the Allen Institute for Artificial Intelligence, those ethical principles stated by big corporations and research laboratories are grouped into 12 common principles. The analysis reflects a broad international consensus on at least 10 out of those 12 principles showing a wide homogeneity and common willingness to move into making AI beneficial for humans (Etzioni & Decario 2019).

Even if the ethical principles above have been, in general, vaguely defined and, in some cases, reflect some aspirational long-term humanity-level objectives, the community had a basis and was ready for the creation of international policy guidelines, at governmental level, that could regulate the AI systems.

In this context, the World Economic Forum's Centre for the Fourth Industrial Revolution, developed jointly with the UK government guidelines that "would help governments make informed procurement decisions and allow both established and new AI providers to compete on a level playing field for government contracts" (WEF 2019). The work started then has been continuously updated to consider all new factors that need to be taken into account when acquiring this type of technology. The recommendations include considering the relevant existing legislation (to defend the rights of citizens), highlight the technical and ethical risks and incorporate mechanisms of accountability and transparency.

In 2019, the AI Group of experts at the Organization for Economic Co-operation and Development (OECD) including over 50 experts from different disciplines and different sectors (government, industry, civil society, trade unions, the technical community and academia), published the OECD Principles on AI in order to promote an AI that is "innovative and trustworthy and that respects human rights and democratic values" (OECD 2019). The text was adopted in May 2019 by OECD member countries (38 countries as of 2022, including USA, the EU, the UK, Korea and Japan).

The OECD AI Principles are the first such principles signed up by governments. They include concrete recommendations for public policy and strategy, and their general scope ensures they can be applied to AI developments around the world. An AI Policy Observatory was consequently created in order to support governments in the implementation of those principles and share best practices.

The OECD AI Principles are:

- Inclusive growth, sustainable development, and well-being
- Human-centred values and fairness
- Transparency and explainability
- Robustness, security and safety
- Accountability

The OECD document introduced the concept of AI system lifecycle reckoning that the ethical principles should be applied in all phases of the design, development, implementation and when using the AI system.

These principles highlighted that the usage of AI should be beneficial for humans and the planet, be inclusive and should contribute to reducing all sorts of inequalities. The OECD insisted on the democratic edge of its plea urging the AI actors to "respect the rule of law, human rights and democratic values", such as "freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognised labour rights" (OECD 2019).

As it will be discussed further later, beyond the need to have AI systems which decision process could be well understood (explainability) and to know when and where those systems are leveraged (transparency), the OECD principles reinforced the requirements for the AI actors to enable traceability of the AI decisions and, when necessary, be held accountable of any malfunctioning or disrespect of the above principles by the system.

The OECD encouraged the development of policies and regulatory frameworks to boost AI innovation respecting the settled principles.

## 1.2. From Guidelines to Legislation

In 2020, the Council of Europe's Ad-hoc Committee on Artificial Intelligence (CAHAI), insisted that recommendations and guidelines are not enough to regulate the development of the AI systems. While acknowledging that "soft law" instruments issued by governmental and nongovernmental organisations (including private companies and academic organisations) are useful tools, "soft law approaches should not be considered substitutive of mandatory governance. Due to conflict of interest, self-regulation efforts by private AI actors are at particular risk of being promoted to bypass or obviate

mandatory governance by governmental and intergovernmental authorities" (CAHAI 2020).

The "soft law" approaches, as noted by the Alan Turing Institute, are "non-binding and rely on voluntary compliance which can lead to varied practices and outcomes" or result into cosmetic commitments to ethical AI (ATI 2021).

Indeed, a step further seemed necessary to steering the development of AI systems for social good and in abidance of ethical values and legal norms. This process had to be done transparently to inform the public on the approach used and avoid misunderstandings regarding rights and freedoms. Moreover, the policy should be homogeneous across the countries in order to avoid any competitive advantage that could be gained by countries tempted to lower the regulatory threshold of AI companies.

The analysis of the CAHAI showed that "the existing framework based on human rights, democracy and the rule of law can provide an appropriate and common context for the elaboration of a more specific binding instrument to regulate AI in line with the principles and values enshrined in the international legal instruments, capable of addressing more effectively the issues raised by AI."

One of the major identified threats to avoid is that AI systems could dehumanise individuals through machine-driven tools and solutions. The respect of human dignity should be paramount all along the way.

Finally in April 2021, after several consultative rounds, the European Commission issued the proposal for an EU AI Act (EC 2021). It was the first (proposed) law on AI set by any major regulator anywhere. The legal basis for the proposal is in the Treaty on the Functioning of the European Union (TFEU), which provides for the adoption of measures to ensure the establishment and functioning of the internal market.

The regulation follows a risk-based approach, differentiating between uses of AI that create (i) an unacceptable risk (prohibited practices), (ii) a high risk (regulated practices), and (iii) low or minimal risk (unregulated).

The list of prohibited practices covers those that pose a significant potential to manipulate persons (including certain type of profiling or using deep fakes), exploit vulnerable groups (like children), violate existing data and consumer protection laws, or violate the digital service legislation. A particular mention is given to forbid the public authorities to perform social scoring of individuals and use, except in limited exceptions, real-time remote biometric identification systems in publicly accessible spaces for the purpose of law enforcement (these two practices are allowed in China for instance). Note that "scoring" at social and judicial level could lead to AI predicting the potential behaviour of individuals in the future and therefore condemning them before having committed any offense (as in the dystopian reality described in Philip K. Dick's "The Minority Report").

AI systems in the second category are those that create a high risk to the health and safety or fundamental rights of natural persons and should follow mandatory requirements and an ex-ante conformity assessment. These concern applications related to management of critical infrastructure, employment and worker management, educational trainings,

biometric identification, law enforcement, administration of justice and democratic processes. The requirements to "high-risk" AI systems include legal obligations in relation to:

- Risk Management
- Data and data governance
- Technical Documentation
- Record-Keeping (traceability)
- Transparency and provision of information to users
- Human oversight
- Accuracy, robustness, and cybersecurity

At Union level, the proposal establishes a European Artificial Intelligence Board (EAIB) to facilitate harmonised implementation of the new rules and to ensure cooperation between the national supervisory authorities and the Commission. At national level, Member States would have to designate a national supervisory authority.

Unfortunately, as probably expected, Article 2 of the Act states that "this Regulation shall not apply to AI systems developed or used exclusively for military purposes." Those would be the exclusive remit of the Common Foreign and Security Policy regulated under Title V of the Treaty on the European Union (TEU).

Note that on the point above regarding biometric identification, Andrew Moore director of Google Cloud AI and commissioner of the US National Security Commission on

Artificial Intelligence, highlighted the Chinese government's "Orwellian" use of surveillance technologies (Jasper 2021). Already some prominent companies have taken a stance on facial recognition and IBM's CEO Arvind Krishna announced in 2020 that "IBM firmly opposes and will not condone uses of any technology, including facial recognition technology offered by other vendors, for mass surveillance, racial profiling, violations of basic human rights and freedoms, or any purpose which is not consistent with our values and Principles of Trust and Transparency" (Krishna 2020). The EU Act tries to restrict and regulate its usage (already regulated under the General Data Protection Regulation or GDPR) with strong requirements (high-risk category) and forbids its usage for law enforcement purposes except if few exceptions and under judiciary control. The debate is far to be closed regarding this very sensitive item.

Even though this important step forward, and as noted by the Alan Turing Institute, the current legal mechanisms protect individual rights but still fail to thoroughly address risks at the societal level like those associated with electoral processes or democratic institutions (ATI 2021). Further public oversight and involvement is necessary to protect the democracy, and this comes with either modernising existing binding legal instruments or adopting a new binding legal one. No-biding approaches will fail to fulfil this governance task.

The proposal for an EU AI Act follows now procedural steps (amendments are possible) and is expected to be approved within two years. This initiative corroborates the EU interest in being a normative leader and values exporter in anything that concerns the Digital world. EU's GDPR is already considered gold-standard across the world.

Having the EU defining the standards on AI has raised some eyebrows across the Atlantic and some claim that the US should collaborate more with the EU in order to avoid AI applications become ‘too restrictive’ (Gaumond 2021).

## 1.3. AI Ethics in China

In 2017, China’s State Council set out an ambitious plan to leapfrog the US to become the global leader in AI. In February 2019, the China National Governance Committee for the New Generation Artificial Intelligence was established under the Ministry of Science and Technology of the People's Republic of China (MOST). In September 2021, just 2 months before the global UNESCO Recommendation that will be described just after, the Committee issued its New Generation of AI Ethics Code (IRC 2021) which plaids to integrate ethics into the full life cycle of AI and is to be applicable in China immediately.

The stated goals are: “to enhance the ethical awareness on AI and the behavioural awareness of the entire society, to actively guide the responsible AI research, development, and application activities, and to promote healthy development of AI”. The scope of application of the norms includes natural persons, legal persons, and other related organizations engaged in related activities such as management, research and development, supply, and use of AI.

The ethical norms defined in the released document are:

- Enhance the well-being of humankind.
- Promote social fairness and justice.
- Protect privacy and security.
- Ensure controllability and trustworthiness.
- Strengthen accountability.
- Improve ethical literacy.

The first principle refers to the respect of human rights and the fundamental interests of humankind. The description of this principle also includes the need to "adhere to the priority of public interests" which may contradict the respect of individual rights. As we know, this duality is common in China where individual rights can easily be cancelled on the benefit of 'public interest' as seen during the management of Covid-19 in the country.

The description of the first principle also asks AI applications "to promote human-machine harmony, improve people's livelihood, enhance the sense of happiness, promote the sustainable development of economy, society and ecology, and jointly build a human community with a shared future."

The "protecting privacy and security" norm asks to protect personal data and privacy in accordance with the principles of lawfulness, justifiability, necessity, and integrity. Data must not illegally collected and the rights of personal privacy respected.

As reported by the South China Morning Post, “the emphasis on protecting and empowering users reflects Beijing’s efforts to exercise greater control over the country’s tech sector” (Shen 2021) and follows the efforts done by the government to regulate the use of content recommendation algorithms, often based on AI. The Chinese internet’s watchdog, in a document also signed by Propaganda Department of the Communist Party, recalled that the inappropriate application of algorithms poses “a challenge to the protection of ideology, social justice and the rights of internet users” (Shen & Qu 2021). Again, the way that China controls the internet and manipulates the information delivered to its citizens is illustrative on the duality between the principle formulation and the reality, at least from the perspective of democratic countries.

The fourth principle is very interesting and states that humans should “have the right to choose whether to accept the services provided by AI, the right to withdraw from the interaction with AI at any time”. It can be related to the right to understand (transparency) and challenge the decisions done by AI systems as proposed in the EU AI Act.

This forth principle also covers the need to have “human oversight” (control) on any AI system allowing the humans to suspend its operation if needed. In this sense, as declared by the last principle, humans should be the liable subjects (responsibility) of any harm produced by AI systems.

The document, beyond these six basic principles, expands the normative requirements to the areas of management, research and development, supply and use of AI systems.

The Norms for Management warns for instance about the temptation of "rushing for quick success" with shortcuts in the requirement to have healthy and sustainable development environments. Management should not abuse of power, needs to protect privacy and dignity of all stakeholders, and promote inclusivity and openness. It is interesting to see that it seems necessary for the Committee to recall those basic management ethical rules that should be applied to any sector, and not only the AI one.

The Norms of Use recall the need to do risks analysis and mitigation prevention before launching any AI system in order to avoid malicious use, misuse and abuse.

The caveat of setting ethical principles in authoritarian regimes is that they are subject to arbitrary laws and these laws to the regime or the ideology. The norms described in this chapter could be fully reused by any democratic system but their application in China, taking into account its track record of human rights abuse and manipulation, is yet to be proved. They may show, however, the willingness of China to be accepted as a rightful partner in this technology race.

## 1.4. The First Global Recommendation on the Ethics of AI

In 2019, UNESCO embarked on a two-year process to elaborate this first global standard approaching this task from a multidisciplinary perspective involving a wide range of stakeholders. The main driver was the conclusion of UNESCO's World Commission on the Ethics of Scientific Knowledge and Technology (COMEST) that emphasized that "at the global level, there is a need for a general universal ethical guidance in terms of core values that must underpin the development of AI systems." It recommended UNESCO to

leverage its normative mandate to raise the awareness about the ethical impact of AI on the various social, cultural and scientific aspects of society (UNESCO 2019).

On 24 November 2021, the Recommendation on the Ethics of Artificial Intelligence was adopted by UNESCO's General Conference at its 41st session.

The scope of the Recommendation addresses ethical issues "in relation to the central domains of UNESCO: education, science, culture, and communication and information" and is addressed to Member States, both as AI actors and as authorities responsible for developing legal and regulatory frameworks. It also provides ethical guidance to all AI actors, including the public and private sectors, "by providing a basis for an ethical impact assessment of AI systems throughout their life cycle" (UNESCO 2021).

UNESCO's Recommendation focus not only the States. The Recommendation provides principles and values for sound policy making around AI technologies, but it also aims at guiding the actions of individuals, institutions and private sector companies that develop and apply them. The involvement of the private sector is key as major technology companies lead this area.

The end goals of the UNESCO's Recommendation are:

> "*to protect, promote and respect human rights and fundamental freedoms, human dignity and equality, including gender equality; to safeguard the interests of present and future generations; to preserve the environment, biodiversity and ecosystems; and to respect cultural diversity in all stages of the AI system life cycle*"

There will be contradictions while applying properly the values recommended by UNESCO. For instance, it might be considered that a face recognition system based on

AI may help protecting at least some of the citizens' rights (right to security). However, its implementation may contradict others like introducing gender bias or go against other fundamental freedoms.

In those cases, it is asked that any limitation to human rights and fundamental freedoms has to be based on a law ('lawful'), be reasonable and proportionate, and consistent with International Law.

It is important to note that the respect to the principles of AI has to encompass its entire 'system life cycle'. Therefore, it is not enough to use an AI system respecting the fundamental freedoms. Its conception, design and deployment have to also adhere to those same principles. A facial recognition system can be lawfully used in some very controlled cases but if during the machine learning phase of the model, the designers have not leveraged samples and data sets that avoid gender and racial discrimination, when using the system, the results will be discriminatory or biased, even if that was never the intention of the end user.

As reminded by the CAHAI (2020), "AI-applications currently being used could enshrine, exacerbate and amplify the impact on human rights, democracy and the rule of law at scale, affecting larger parts of society and more people at the same time." A well-known example of this when Microsoft engineers created an AI system that could imitate a human sending tweets. The bot learned from the users that interacted with it and soon became a machine sending racist and misogynist tweets. Microsoft said: "the AI chatbot Tay is a machine learning project, designed for human engagement. As it learns, some of its responses are inappropriate and indicative of the types of interactions some people are

having with it" (Hunt 2016). The machines will learn from the society and just replicate its behaviours. It is up to the different AI actors to make sure that the "good" side of a society is leveraged into the learning process of those systems.

The values that the UNESCO Recommendation on AI put forward are:

- Respect, protection and promotion of human rights and fundamental freedoms and human dignity
- Environment and ecosystem flourishing
- Ensuring diversity and inclusiveness
- Living in peaceful, just and interconnected societies

It is explicitly written that "new technologies need to provide new means to advocate, defend and exercise human rights and not to infringe them". The AI systems should be conceived then to enhance humans' quality of life and not the contrary. This encompasses the need for AI actors to play an enabling role to ensure peaceful and just societies (UNESCO 2021).

These values are translated into a more detailed set of principles that AI systems should embrace:

- Proportionality and Do No Harm
- Safety and security
- Fairness and non-discrimination

- Sustainability
- Right to Privacy, and Data Protection
- Human oversight and determination
- Transparency and explainability
- Responsibility and accountability
- Awareness and literacy
- Multi-stakeholder and adaptive governance and collaboration

Even if the 193 UNESCO's state members signed this Recommendation, military applications are not within the scope of it. Expectedly, military applications using AI will not be conceived to "enhance" human quality of life and provide no harm.

However, two of the above principles, "human oversight and determination" and "responsibility and accountability" are the centre of the debate on autonomous military AI systems as we will see afterwards. A specific treaty on AI-based arms control could be justified building upon the need to respect these principles.

The first one states that states "should ensure that it is always possible to attribute ethical and legal responsibility for any stage of the life cycle of AI systems […] to physical persons or to existing legal entities." While the AI systems are supposed to make decisions on their own, they can do so because a human has decided it, has activated the system, and has determined which outcomes the AI system can provide and which consequences those outcomes can yield. This human responsibility has to be understood by the user of the AI system who can be held accountable of any harm produced

purposedly or un-purposedly by it. The UNESCO recommendation pushes the states to make legally liable the AI actors.

The second key principle cited above, “responsibility and accountability” reinforces the concept of liability as the AI actors should assume the ethical and legal responsibility of AI systems in accordance with national and international law. To hold anyone accountable means that the state or court should be able to clearly establish the responsibilities of each one of the actors involved in the AI system. This requires transparency in the way the AI systems are designed, managed, and controlled. To guarantee this transparency, the end-to-end phases of the system should be allowed to be audited if necessary and the responsibilities traced back to humans or legal entities as the previous principle required.

The “safety and security” principle states that “unwanted harms (safety risks), as well as vulnerabilities to attack (security risks) should be avoided and should be addressed, prevented, and eliminated”. We will see later that this principle is similar to the “reliability” principle used in the NATO recommendation which adds ‘robustness’ to safety and security.

Establishing confidence and transparency in AI systems and promoting democratic values do not just apply to the civil world and of course concern the states’ security, at least for the democratic ones. The US National Security Commission on Artificial Intelligence for instance called the government to “focus on ensuring that its AI systems are robust and reliable” and that government use of AI is “effective, legitimate, and lawful”. This in turns pushes for the need to develop AI tools to “enhance oversight and auditing,

increasing public transparency about AI use, and building AI systems that advance the goals of privacy preservation and fairness" (NSCAI 2021).

The debate around AI ethical values has just recently moved into geopolitical spheres. The US National Security Commission on Artificial Intelligence highlighted that the "AI competition is also a values competition. China's domestic use of AI is a chilling precedent for anyone around the world who cherishes individual liberty. Its employment of AI as a tool of repression and surveillance—at home and, increasingly, abroad—is a powerful counterpoint to how we believe AI should be used" (NSCAI 2021). This sets the tone into further competition on AI between both countries but also is a chilling warning about how authoritarian regimes can use AI and on how unprepared can democracies potentially be to counterbalance malicious AI tools if they do not research and develop those same tools.

As for any "dual use" technology, like the Internet, that can be used both for civilian and military purposes, AI ethical recommendations overlap with security concerns either because the state feels the need to keep the control of AI capabilities or because the "wrong" usage of AI can threaten state values.

# CHAPTER 2: AI and Global Security

## 2.1. AI and Security Threats

In July 2017, China laid out plans to become the world leader in AI by 2030 (Kharpal 2021). China leads already the race of quantum computing and claimed in 2020 to have a system that surpassed Google's 2019 achievement by a factor of 10 billion, according to the Xinhua News Agency (Guterl 2020). Guterl's article title is illustrative: "As China Leads Quantum Computing Race, U.S. Spies Plan for a World with Fewer Secrets", quantum indeed will make current cryptography techniques useless.

Even if progress in quantum computing is mostly experimental today, the Chinese goals and levels of investment on AI and quantum computing have been a waking up call for Western democracies. The words of Putin in 2017 resonate well: "[AI] comes with colossal opportunities, but also threats that are difficult to predict. Whoever becomes the leader in this sphere will become the ruler of the world" (Vincent 2017). Few months later, Eric Schmidt, the former CEO of Google, argued that the US was facing its "Sputnik" moment indicating that the AI is very important to the future of power and that the US needs a national strategy on AI, as it had one for the development of space technology during the Cold War (Clark 2017).

As reminded by Meserole (2018), "new technologies introduce uncertainty about military capabilities: each advance brings with it uncertainty about how it will be used, or even how powerful it will be". In the case of AI, this sparks the question on how advanced my adversary is, but yet, advanced with regards to what and on which capabilities? AI is

indeed an enabling technology, not a weapon by itself. AI can be embedded in multiple types of weaponry, convert traditional tanks, submarines or planes (drones) into semi or fully autonomous systems and replace human decision making in command and control systems.

AI technology also raise uncertainties on how adversaries can effectively use it during a conflict, are there new weapons or strategies that I do not know yet that they exist? Would AI change the way conflicts are built up and fought? Finally, we do not know yet how harmful a fully AI-driven military can become. An AI-driven war can have worse consequences for me than a nuclear one? Is there any deterrence possible and if yes, how can I build it and make sure I am in leading it?

This leads to a classical security dilemma where uncertainty on the intentions and capability of the adversary regarding AI military applications provokes enhanced level of investment in military spending. This in turn can be considered by the adversary as a threat consequently making him raise its own spending.

An AI arms race seems well on its way today. The Future of Defense Task Force report 2020 advances that "AI-enabled weaponry driven by speed and precision compete in a complex battlespace, [and] requires the US to invest significantly in both offensive and defensive AI capabilities" (HASC 2020). The conclusion of the US National Security Commission on Artificial Intelligence is also clear: "the US must act now to […] invest substantially more resources in AI innovation to protect its security, promote its prosperity, and safeguard the future of democracy. […] The government should double non-defense funding for AI R&D annually to reach $32 billion per year by 2026" (NSCAI

2021). The Commission chair is no other than Eric Schmidt. Note that leadership in AI would mean general economic leadership and consequently capacity to fund the military.

This technological (and arms) race also has a moral driver. The language used might be just the right way to awake the sensibilities of the lawmakers and motivate them to approve the budgets but is significative how the technology itself becomes here the actor of the moral debate. NSCAI Commissioners emphasized that their recommendations are meant to "instil" AI and emerging technologies with "American values." The report devotes a chapter to upholding democratic values that calls for greater oversight and transparency around how the government uses AI. Instead of building, as done until today, neutral arms that will help states to defend the "American values", AI has to be built (and consequently the AI weapons too) on top of those values and by doing so achieve the defence of the democracy and human dignity.

Establishing confidence and democratic models for AI are key objectives. The NSCAI Commission urges the government to defend against emerging AI-enabled threats to America's free and open society, prepare for future warfare, manage risks associated with AI-enabled and autonomous weapons and transform national intelligence (NSCAI 2021). The tone is set for a major overhaul of the security structures in terms of military capabilities.

Beyond the security dilemma, AI technology will change the configuration and profiles of the armies, set new standards of military superiority beyond brute force and consequently challenge, at least temporarily, the current balance of power (Horowitz 2018).

In a battlefield, with myriad of sensors and cameras of all sorts deployed, the amount of information generated will be massive and no human, only high computational AI systems, could deal with it. The battlefield in the end is no more than a game grid. Every asset is traced and calculations on correlation of forces are made deciding which actions to be taken based on a risk-benefit analysis and efficient usage of the force to inflict the highest possible damage. All decisions could be made automatically and towards autonomous tanks, drones or other weaponry which feedback in real time new data, results, or changes on the field. This, intrinsically, is not much different from what a reinforcement learning algorithm can already do. An example is AlphaGo, a software that beat the human world champion in a game that has more possible positions than atoms in the Universe. The algorithm played against himself, millions of times to refine its knowledge of the game and, during the game itself, it showed a level of 'creativity' that even surprised expert players.

Other examples of how AI is shaping already current developments and conflicts are (Husain 2021):

- In 2020, the Defence Advanced Research Projects Agency (DARPA) organized a dogfight competition between human F-16 pilots and various AI algorithms, called "AlphaDogfight". Even if some critics considered that the competition was not fully fair, AI won 5-1.

- During the war in Syria, the Russian army used 80 Unmanned Aerial Vehicles simultaneously over the battlefield. The Russian Defence Minister Sergei Shoigu commented that the experience was like a “semi-fantastic film”.

- During the Azerbaijan-Armenia conflict, Turkish TB2 drones were massively deployed against conventional forces leaning decisively the war towards the Azerbaijani side.

- Azerbaijan has converted old planes into DEAD (Destruction of Enemy Air Defence) drones by using them to both identify Aire Defence sites and destroy them via kamikaze attacks. China has converted his old J-6 and J-7 aircraft into AI autonomous drones.

- Russian military is designing autonomous vehicles to guard its ballistic missile bases as well as an autonomous submarine that could carry nuclear weapons (Bendett 2017).

AI systems would enable massive coordination and optimization of force. As a result, a small, highly mobile force (e.g. drones) under the control of AI could always outmanoeuvre a much larger conventional force.

A concerning security threat is that the barrier of entry for new players to get access to his new technology and weaponry has been lowered significantly.

Turkey is considered a leading player in autonomous drones driving most of the innovation in this area. Yet, its defence spending is 70 times smaller than the US one (GFP 2022). Turkey's TB2 drones are 60 times less expensive than a single US attack helicopter, and this does not count the fact that the helicopter needs two highly trained soldiers to be operational (beyond the fact that they obviously risk their life operating it), the helicopter needs much more fuel to be on air and its transportation to the battlefield is cumbersome. Eight of those TB2 drones would have the same fire power than such helicopter (Husain 2021).

Turkey's lethal autonomous weapons systems (LAWS) deployed in Libya are credited to have been the first ones to hunt down and kill humans: "spotted live targets using AI and destroyed them with an autonomous strike based on the information in the database, without the need for any commands" (Hernandez 2021). Other type of drones can carry lasers that can target and shut down other drones.

As seen above, the competitive advantage of large economies and huge armies is reduced with this new technology. The access to the technology is easier and the costs lower. Poorer nations and terrorists can now afford to have challenging armed forces. As demonstrated in the 2020 Azerbaijan-Armenia war, few drones can now defeat a conventional army.

Countries will start to phase out tanks as a viable military platform, as already understood by the US Marine Corps (Snow 2020). The configuration of military spending is already shifting, and the configuration of the armies will follow. This creates a worrisome situation for the major powers, that can see their military advantage eroded.

As pointed by Detsch (2021), “it’s becoming easier to hunt and kill troops [on the ground] than ever before, and to do so on the cheap” and, even more dooming, drones are increasingly harder to take down.

All the above items just show what AI can already do now. We can hardly imagine what could come next. AI systems could be used in the future in sophisticated arms programs (e.g. hypersonic, space) and in next generation cyber and information warfare.

## 2.2. AI and Cyberwar

The Internet is now a global network of networks of private, public, academic, business, and government entities linked by a broad array of electronic, wireless, and optical networking technologies. Even if it is not yet a universal asset, its centricity in today’s world is nowadays patent on how it shapes economic transactions, work behaviours, financial exchanges, information flows and communications, among others.

The Internet, or the cyberspace, is a key element of our economic, social, political, and organizational life, and therefore, subject to crime and conflict. Terms like cyberconflict, cybercrime, cyberattack, cyberespionage, cyberterrorism or cyberwar have recently become common, translating how the technical capabilities of governments, organised groups, or individuals (hackers) are now able to deeply disrupt the Internet in the pursue of selfish and evil purposes.

In a rapidly evolving technological and integrated society, where all things and systems are connected, the consequences of this type of attacks can be more and more important.

Attacks that in the past mostly focused on espionage and intellectual property theft have evolved into creating higher disruptions impacting the information systems in health care, transportation, businesses, and governments. The attacks can now lead into the degradation of industrial infrastructure or provoke accidents that affect people. In the future, criminals could hack self-driving cars, auto-piloted planes, and drones, or even AI killer robots.

The cyberspace has become a new field of confrontation between rival states. Its novelty, its span and its intrinsic technological nature have raised concerns on whether regulation would be necessary at least to avoid devastating collateral damage to innocent population.

The threats are very real with states using already AI systems to enhance disinformation campaigns and cyberattacks. The collection of huge amounts of data from the citizens can later trigger targeted attacks with the objective of manipulating, coerce or blackmail individuals or entire societies. As recommended by the US National Security Commission on Artificial Intelligence, “the government should leverage AI-enabled cyber defences to protect against AI-enabled cyberattacks. And biosecurity must become a top-tier priority in national security policy.”

Also note that the revolution of quantum computing will sum another notch to the level of uncertainty and risk associated with the cyberspace and the cyberattacks as quantum will literally convert into non relevant most of the current encryption based and security methods used worldwide.

Cyberattacks not only come from adversary states. Criminal organizations or activist groups like Anonymous can easily disturb the major internet services of a country. Governments can even hire freelance hackers to create disturbance to another country or, as it happened with Qatar in 2017, flood the official news agency with fake news where the country leaders verbally attack allies, in this case President Trump, or praise antagonist terrorist groups (Kirkpatrick & Frenkel 2017).

This illustrates how hard is to attribute attacks in the cyberspace. Hidden identities, despite recent efforts in this sense, are hard to uncover and, when the origin of the attack is somehow found, it is hard to prove and verify the veracity of the information without uncovering the technology used, and therefore be exposed to new weaknesses. On top of that, discovering that the attack comes from a research lab, does not necessarily prove that the research lab is responsible of it as the hacker can even hide somewhere else.

The security dilemma in the cyberspace can have a more vicious twist. As exposed in Riordan (2018), a country that fears a cyberattack from another one may decide to penetrate the systems of the adversary to understand its intentions and capabilities. The adversary can detect the attack and interpret it as being a first step for a larger scale one with maybe worse malicious intentions. In the cybersecurity dilemma is not possible to distinguish between defensive and offensive actions and any attack can be the first step for a more disruptive one. Attacks which purpose could be disruption, degradation or just espionage can look be very similar.

Deterrence in cyberspace is therefore much difficult to put in place. Note also that the countries that would suffer the most of a cyberattack are the most advanced ones where

civil societies, industries, devices, and services are fully interconnected and computerized. A cyberattack towards countries like North Korea would have little impact and this asymmetry with regards, for instance, the US would make hard to set a deterrence uniquely based on reciprocal attacks between these two countries. Riordan (2018) remarks that China has for instance changed its cybersecurity strategy, becoming less aggressive, as the country became more interconnected and its economy fully integrated. "This suggests that informal modes of restrained behaviour have already been employed for at least the last fifteen years [between the major powers]" (Valeriano & Maness 2018).

Nye mentions normative taboos (or ethical approaches) as a relevant mean for deterrence promoting the establishment of a taboo not against types of weapons but against certain types of targets (Nye 2017). As he writes, "the US has promoted the view that the internationally recognized laws of armed conflict (LOAC), which prohibit deliberate attacks on civilians, apply in cyberspace". The internet is as a "dual use" system used for civilian and military purposes and distinguishing between a military or a civilian target may prove challenging.

A key issue is also identifying how to respond to a cyberattack. The Tallinn Manual, adopted by NATO, allows for the possibility of responses in both physical and cyberspace (Riordan 2018). Within this spirit the US government has made clear that they will "respond to cyberattacks 'in the time, manner, and place of our choosing'. This is the language of restraint, not deterrence. Deterrence means assured response such that the attacker figures the costs of an attack are too high" (Valeriano & Maness 2018).

AI in the cyberspace can be leveraged to increase the scale and effectiveness of social engineering attacks. The analysis of the citizens data can lead to detect patterns that allow intruders to manipulate them exacerbating extremist positions or distributing targeted fake news.

More classical cyberattacks can become more harmful thanks to AI as they could adapt in real time to defensive approaches and find faster the breaking points of the system. AI attacks could navigate faster within the penetrated systems and produce more harmful damages. In a reciprocal way, a cyber defence based on AI could create cyber resiliency identifying quickly disguised attacks and reacting consequently.

## 2.3. AI and LAWS

A special place in the international debate has been taken by the recent development of LAWS: Lethal Autonomous Weapon Systems, in some cases referred as Killer Robots. As the world could witness in Libya, these weapons belong to the present and not a distant future. The development of robotics and AI will increase their precision and lethality. Their low cost will make possible to use them in swarms that can overwhelm any current defensive systems that are not prepared to address large number of small and fast targets. Specific electronic warfare will need to be developed to stop them but that could also hurt the humans we try to protect. In the case of very small LAWS that could carry poison or tiny explosives that could kill a human, stopping them could be impossible.

In a not too far scenario, we could imagine showing a picture of someone to a LAWS and let them roam, find the target and kill it. Authoritarian regimes could even launch ethnic genocides through them (Tegmark 2017).

As noted in the 2020 Future of Defense Task Force Report, “it is imperative that policy experts and lawmakers consider the second- and third-order effects of developing and deploying LAWS. Moral, ethical, and legal factors will need to be weighed accordingly” (HASC 2020).

The UN Convention of Certain Conventional Weapons (CCW) Group of Experts, established in 2019 eleven principles governing LAWS (CCW 2019). Among them, the CCW made clear that International Humanitarian Law (IHL) fully apply to them and that human responsibility for decisions on the use of weapons systems must be retained since accountability cannot be transferred to machines. The CCW recalls that it offers the appropriate framework for dealing with new weapons issued from the application of new technology (like AI) with the purpose of find a balance between military necessity and humanitarian considerations.

IHL prohibits any (autonomous) weapon that (a) has characteristics prohibited by a weapon treaty or customary law; (b) is of a nature to cause superfluous injury or unnecessary suffering; (c) is indiscriminate by nature; or (d) is intended, or may be expected, to cause widespread, long-term, and severe damage to the natural environment.

An important caveat is highlighted by the Stockholm International Peace Research Institute (SIPRI), “the respect for IHL presupposes the ability to foresee, administer and trace the operation, performance, and effects of AWS” (Boulanin et al. 2021). It must be possible to understand if the weapon does not respect the established conventions and prohibitions. On other words, the weapon behaviour needs to be, at least at some level of

degree, predictable and transparent. AWS governance during the conflict must also satisfy the rules governing the conduct of hostilities according to IHL and respect the principles of distinction, proportionality, and precautions. Consequently, when used, AWS effects should be able to be traced and analysed in order to establish criminal responsibilities in case of non-respect of the IHL.

As the CCW stated, human-machine interaction should ensure that the potential use of LAWS is in compliance with applicable international law, in particular IHL. However, the level of autonomy of a LAWS can vary greatly, have several levels of degrees, apply to different operational components of the system or just be triggered in case of very specific circumstances. Countries agree that further clarification is needed on the type and degree of human-machine interaction required, including elements of control and judgement, in different stages of a weapon's life cycle, in order to ensure compliance with IHL (CCW 2019). The Pentagon guidance requires that "appropriate levels of human judgment" preside over the use of force by LAWS (HASC 2020), although what is appropriate is open to broad interpretation.

SIPRI makes a call for further international cooperation in this area. States need to develop standards and ethical bases regulating the expected behaviour of LAWS to fully comply with IHL including the possibility to be able to trace back the responsibilities of illegal use of LAWS to be able to prosecute individuals or states.

According to a 2020 Human Rights Watch report, "China, Israel, Russia, South Korea, the UK, and the US are investing heavily in the development of various autonomous weapons systems, while Australia, Turkey, and other countries are also making

investments” (HRW 2020). Since 2013, only 30 countries have called for a ban on fully autonomous weapons.

Some arguments in favour of the usage of LAWS include that by being able to better detect and target military troops or other LAWS, they would allow to reduce the number of civilians casualties.

LAWS could also lead to have a reduced number of human soldiers deployed in the battlefield reducing the risk of loss of life, even if this argument would only apply to the army that owns the LAWS, i.e. the most advanced countries. A counterargument to this point would be using LAWS could potentially multiply the number of armed conflicts as less troops will need to be mobilized and hence decreasing the human cost of conflicts.

Some scientists like Ronald C. Arkin believe that autonomous robots can actually act more “humanely” (Arkin 2015). For instance, they do not need to protect themselves or have a self-preservation imperative and therefore “there is no need for a 'shoot first, ask-questions later' approach, but rather a 'first-do-no-harm' strategy can be utilized instead”. Even if Arkin considers the lack of self-preservation as natural to all machines, it can be interpreted, in reality, as a moral choice of their developers. Indeed, the LAWS owner could consider that a multi-million dollar machine is more valuable than a collateral human life and put self-preservation first and going back to the principle of 'shoot first, ask-questions later'. Note that the moral choice does not belong here to the machine, that simply does what it has been instructed to do, but to the human developer or commander. This illustrates the need for further international alignment and agreements in this area.

Interestingly the fact of lacking “human” features, like the natural instinct of survival, the fact they do not get nervous in stressful situations, the fact they do not get choleric or emotional in front of enemies or the fact that they can process a higher amount of information, would make them act in a much more “human” way in the battlefield.

Arkin also points out that LAWS can also collect precious information in real time regarding the behaviour of all parties during the conflict making possible then to fact checking claims of non-respect of IHL. This could limit the leeway of the parties in conflict pushing them to be more respectful of International Law.

Etzioni highlights the work of military ethicist George Lucas Jr. (Lucas 2013). Lucas points out, for example, that robots cannot feel anger or a desire to seek retaliation. “The debate thus far has been obfuscated by the confusion of machine autonomy with moral autonomy”. For Lucas, the primary concern of engineers and designers developing autonomous weapons systems should not be ethics but rather safety and reliability, which means taking due care to address the possible risks of malfunctions, mistakes, or misuse that autonomous weapons systems will present (Etzioni 2018). However, while the current state-of-the-art technology would probably make unnecessary to define moral considerations for LAWS, it might well be possible that advances in computing science would make necessary to work on them in the near future.

Setting ethical limits to LAWS seems a complex task. But then, as Tegmark points out, if we cannot enforce them to be 100% ethical, why building them in the first place? Wouldn’t that be the easiest option? (Tegmark 2017 p.117).

Surprisingly or not, as seen above in the work of Lucas, some of the proposed ethical values that were defined by the UNESCO Recommendation for civil applications resurface in the military sector. Ethical principles are even requested by the states that lead LAWS technology. UNESCO's values like proportionality, safety, security, human oversight, transparency, explainability, responsibility and accountability seem more and more commonly discussed in military forums and are pushed forward in other to regulate the usage of AI arms.

## 2.4. AI and Hyperwar

We mentioned previously the possibility of imaging a battlefield with myriad of sensors, multiple radar systems and high-definition infrared and conventional cameras of all sorts and sizes generating massive amounts of data. An AI system as command and control of the operations could process at light speed that information and make decisions involving autonomous tanks, drones or other weaponry that can react and move faster than any human controlled vehicle. The results of those decisions being again processed in real time generating new data that would trigger fresh decisions.

The speed at which those decisions could be made at such a big scale make that any human intervention would be almost pointless or come far too late. Obviously, no human could compete against such system.

'Hyperwar' is a term coined already back in 1991 following the First Gulf War whereby tactical dominance in the battlefield is achieved with superiority on the computing side of warfare. War is becoming "unimaginably and unmanageably fast" whereby scientists aspire to delegate decisions to computers (Arnett 1992). Army generals may want to keep

control and make the final decisions but when that would have been an option 30 years ago, it has become impossible today unless they do not exploit fully AI capabilities.

With AI, as pointed by John Allen, hyperwar has gone one level beyond embedding "warfare with highly sophisticated AI algorithms where humans are seldom found in the loop. The role of the human in the loop is and should be the subject of an enormous ethical debate" (Allen 2018). When human decisions are absent from the observe-orient-decide-act (OODA) loop, the time between observation and act will be reduced to near-instantaneous responses. AI algorithms excel at managing huge amounts of data (observe), detect information patterns and key elements to track and identify (orient), make decisions based on states, rules, and scenarios (decide) and, obviously, send the necessary instructions to command actions (act). Data science tools, machine learning, reinforcement learning, artificial vision are technologies ready today.

As stated by the US National Security Commission on Artificial Intelligence, "defending against AI-capable adversaries operating at machine speeds without employing AI is an invitation to disaster. Human operators will not be able to keep up with or defend against AI-enabled cyber or disinformation attacks, drone swarms, or missile attacks without the assistance of AI-enabled machines" (NSCAI 2021).

All the transformative changes explained in this and previous chapters, threaten the established powers and shake their current status. For instance, the before mentioned Commission foresees that US armed forces' competitive military-technical advantage could be lost within the next decade if they do not accelerate the adoption of AI.

This race has been compared with the space race during the Cold War in the 1960's, partially to motivate law makers to approve the needed budgets as seen before but also to signal the transformative impact that AI will have in the industrial and military landscape. Armies will be reorganized, more technology savvy personnel will be hired and weaponry completely rethought. Assumptions that held in the past like the need of heavy, and expensive, weaponry on the battlefield or strong logistical capabilities or deploy a large number of troops, will be challenged with the new AI paradigm.

This transformative model and the potential impact it may have not only in the battlefield by also during the cyberwar at large scale challenges the deterrence models we have seen so far. AI is a technology enabler, a science, that cannot be 'counted' or monitored by specialised personnel like nuclear heads. AI does not need large industrial facilities like the ones needed to enrich uranium or plutonium. AI does not need specific materials or fuel like nuclear ballistic missiles do. AI does not need a nuclear submarine or specialised trucks to be transported.

Deterrence capabilities on highly sophisticated AI weapons are difficult to establish and monitor. Like in the case of cyberwar, making them visible may make them less effective to the country that owns them.

However, the alternative of letting this technology become uncontrollable, unethical, and accessible not only to adversary states but to terrorists or crime gangs, is not the right one. An international AI arms treaty becomes a necessity and, even if it is hard to imagine how it could be architected and what it will cover, it is worth to try.

# CHAPTER 3: An International Treaty for AI

## 3.1. Multilateralism and AI

The international system is now facing issues much different to those existing when the UN was created. Issues like climate change, the focus on Sustainable Development Goals, the societal changes imposed by AI and new types of hybrid conflicts need the active involvement of new actors like the civil society and large corporations. From a diplomatic studies perspective, there is evidence that diplomacy is no longer tied to the idea of a narrow diplomatic corps consisting solely of traditional state diplomats, rather to a larger and more complex diplomatic community of diplomats and non-state actors (Wiseman 2015). Technology has enabled the creation of networks of shared interests and the possibility to communicate key messages easily and instantly to almost every individual. Information filters have been dropped even if this has also unleashed levels of misinformation and manipulation not seen until now.

Non-state actors with capacity to influence the international affairs debate have flourished. As remarked by Constantinou, "anyone from the globalized demos can now become diplomat or an activist-diplomat without much difficulty in view of changes in communication and traveling" (Constantinou 2013). This creates new ordeals to diplomats that need to be able to manage the challenge of communicating with a wider range of counterparts from businesses to Non-Governmental Organizations, think tanks, and the media at the speed of social networks and against fake information. Ruggie (2015) talks of polycentric global governance with the co-existence of public law, companies' corporate governance and civil governance mechanisms for positive change. Indeed, the

moral responsibility that comes with using AI systems is increasingly distributed among regulators, developers, users, and hackers.

As former UN Secretary-General Kofi Annan said, “diplomacy has expanded its remit, moving far beyond bilateral political relations between states into a multilateral, multi-faceted enterprise encompassing almost every realm of human endeavour” (Mahbubani 2013). The increase of the non-state actors, their empowerment, their influence into the opinion of millions across the boundaries, have profoundly and positively changed the profile of the multilateral diplomacy.

AI used in a large variety of technologies will dramatically change the society, as described for instance in the report “For a Meaningful Artificial Intelligence” (Villani et al. 2018), and pose unprecedented threats to democracies and global security. As it is the case of climate change, AI breakthroughs, if not addressed properly could become an existential risk to humanity and tension the multilateral system.

As recalled by Pauwels (2018), “the ability of AI to nudge and control private human behaviour and impact self-determination could increasingly limit the capacity of the UN to monitor and protect against human rights violations”. This capacity is further limited when the private sector and powerful states own the required data and algorithms. These big corporations and powerful states may resist the establishment of a global governance for AI and yet this proves the need to have a strong multifaceted multilateral system able to achieve that goal.

Douglas Frantz, former deputy secretary general of the OECD warns that the dominance of China and the US and a few tech giants creates the real prospect of "digital feudalism [whereby] huge amounts of wealth and power is concentrated in the hands of a few [while] many could be left behind" (Frantz 2018).

The UN can play an active role in animating forums for a truly cooperation between technology firms, state actors and civil society. This effort should run in parallel with the UN traditional normative mandate to create standards (and treaties) and the capacity to monitor them.

The "AI for Good" is a UN platform for dialogue on AI led by the International Telecommunication Union (ITU) in partnership with sister UN agencies. It connects AI innovators with public and private sector decision-makers, and contributes to formulate global strategies to ensure trusted, safe, and inclusive development of AI technologies. Started in 2017, its 2018 report had 66 pages of UN AI activities from 27 UN agencies. In 2021 the report had 230 pages from 40 agencies signalling the increased role played by AI in the development of UN projects (ITU 2021).

The UN Interregional Crime and Justice Research Institute (UNICRI) established in 2015 the Centre for Artificial Intelligence and Robotics to focus on AI and robotics in the context of crime prevention and criminal justice, cyber-crimes and strengthening of international criminal law.

The fact that the UN can lead the international normative effort around AI has been shown by the UNESCO and its work on the Recommendation on AI Ethics as seen previously.

The above are all positive signs of the capabilities of the multilateral system. Considering the societal and security risks associated with the implementation of AI systems, it is imperative to move from recommendations towards legally binding instruments also addressing the military usage of AI.

## 3.2. AI Ethics in the Military Domain

As we have seen during in the chapter dedicated to the LAWS, instilling some ethical values to the development and usage of lethal military applications of AI has been a concern and, somehow a requirement, during few years now. Principles like proportionality, safety, security, human oversight, transparency, explainability, responsibility and accountability have been common in the AI ethics debate.

In 2019, scholars and specialists in AI published at the IEEE (the largest technical professional organization dedicated to advancing technology) an article requesting to establish a moratorium on the development of AWS in order to refocus the work into properly defining the guiding principles for human involvement in the use of AWS, ensure the respect of IHL, avoid the proliferation of AWS usage by illicit users (terrorists, rogue states) and develop protocols to mitigate the risks of unintentional escalation due to autonomous systems (Arkin et al. 2019).

The risks associated with escalation seem very real. A machine learning algorithm that can identify a dog or a pineapple in a picture, needs to process millions of images beforehand to learn and get trained. Even when after using a huge amount of data to learn, the algorithm can give some wrong answers. In addition, to run the model to tell us what

is in the picture, we need to input the data in a very specific format. It has been soon understood that providing ‘millions’ of samples of real theatre of operations to military AI systems can be challenging and the exceptional context at which this machines operate would probably lead them to make a higher number of mistakes than machines working in a much more controlled environment.

This reality is the basis for some proposals, that we will detail later, regarding the ban of using AI decision algorithms in the systems that control the launching of nuclear weapons.

The US Department of Defence (DoD), in 2019, issued its “Recommendations on the Ethical Use of AI”. In addition to recalling that the works on AI should abide to the laws of the US, the Law of War, and International Humanitarian Law, five principles were defined to be respected during the development and deployment of its AI systems (DoD 2019):

- Responsible
- Equitable
- Traceable
- Reliable
- Governable

The “Responsible” principle states that “human beings should exercise appropriate levels of judgment and remain responsible for the development, deployment, use, and outcomes of DoD AI systems”. The “Equitable” one asks to avoid unintended biases. The

“Traceable” one is required by International Humanitarian Law and similar to the principles of accountability, responsibility, transparency and explainability.

The “Governable” principle is linked to the necessity to avoid escalations as highlighted before: “AI systems should [… possess] the ability to detect and avoid unintended harm or disruption, and for human or automated disengagement or deactivation of deployed systems that demonstrate unintended escalatory or other behaviour.” I.e. there should be always the possibility to disconnect the machine.

The DoD paper recommends in addition to have “an integrated, iterative development of technology with ethics, law, and policy considerations happening alongside technological development”. Both tracks, ethics and technology development, need to go hand in hand as it is unworkable for one to wait for the other to adapt.

The importance of the private sector in the military domain is also key. As the Future of Defense Task Force report recalls: “to maintain its technological advantage over competitors, the Pentagon must continue to improve its ability to leverage private sector innovation at scale […] recognizing that the private sector, not the government, is now the leader in research and development investment” (HASC 2020).

In October 2021, the defence ministers of the North Atlantic Treaty Organization (NATO) approved the alliance's first Artificial Intelligence Strategy which establishes guidelines for the use of AI in accordance with international law, develop safeguard mechanisms against the threats from malicious use of AI and encourage the development and use of AI in a “responsible manner” (Principles of Responsible Use). The principles

are aimed at providing coherence for both NATO and Allies to enable interoperability (NATO 2021).

The NATO Principles of Responsible Use of AI in Defence are:

- Lawfulness
- Responsibility and Accountability
- Explainability and Traceability
- Reliability
- Governability
- Bias Mitigation

Unsurprisingly, the above principles match those already set by the US DoD except for "lawfulness" that now is explicitly stated as a principle (in the DoD recommendation the respect to the law was part of the document too).

"Lawfulness" means that "AI applications will be developed and used in accordance with national and international law, including international humanitarian law and human rights law, as applicable".

Interestingly the wording used by NATO reminds the one used by UNESCO. Traceability is used instead of transparency, maybe hinting into a stronger legal accountability in military applications.

Instead of UNESCO's "safety and security" principle, NATO prefers to use "reliability" which adds 'robustness' to safety and security and requires all three to be subject to testing and assurance.

## 3.3. Why is a Treaty on AI Needed?

In 2015, Max Tegmark, from the Future of Life Institute, promoted the distribution of an open letter voicing the concerns, within the community of AI & Robotics researchers, of the use of AI in the military field. It was already signed by more of 17,000 researchers in 2017 (Tegmark 2017 p.114) and an excerpt of it is shown below:

> "*Most AI researchers have no interest in building AI weapons and do not want others to tarnish their field by doing so, potentially crating a major public backlash against AI that curtails its future societal benefits. Indeed, chemists and biologists have broadly supported international agreement that have successfully prohibited chemical and biological weapons, just as most physicists supported the treaties banning space based nuclear weapons and blinding laser weapons."*

Tegmark diminished arguments put forward by some researchers on the benefits of using AI-led robots that 'necessarily' should behave better than humans in armed conflicts. Schmitt recalls that "the further removed the warriors are from their acts of war, the more difficult it will be for them to retain the humanitarian spirit that underlies the law of armed conflict" (Schmitt 1998).

As noted by Jones, the multilateral system is failing to catch up with the challenges and threats that these emerging technologies create. Policymakers are ill-prepared to manage crisis situations in cyber, biotechnology, and artificial intelligence (Jones et al. 2019).

There is no international consensus on the application of the 'law of armed conflict' (LOAC) to cyberwarfare for instance. As it happens with the energy grids or the communications systems that serve both military and civilian purposes, disruptions on the internet itself, or through it, will necessarily violate principles like the one of "Military Necessity" and the consequences of the acts are more difficult to evaluate. The fact that the hostile actors can hardly be identified after an attack in the cyberspace, and the reality that they do not usually wear uniforms, make almost impossible to apply the "Distinction" principle. "Perfidy" is difficult to protect when information flows can be easily disguised behind neutral banners. "Proportionality" of cyberattacks is not measured in terms of casualties but in terms of disruption that can carry heavy consequences and kinetic collateral damage, and response. It is therefore hard to contain, control and judge this proportionality. Consequence of these attacks will be amplified with AI. How to address these principles, as were recalled above by Hugues (2010) could be the object of an international treaty.

In 2019, the Centre for International Security and Strategy at Tsinghua University (China), the Brookings Institution (US), the Berggruen Institute (US) and the Minderoo Foundation (Australia) established an 'informal' international dialog on the development and application of AI-enabled military systems.

In December 2020, the Berggruen Institute published an op-ed expressing the belief from US and China participants that an international treaty on the usage of AI in armed conflict is necessary (Allen & Fu 2020). Reckoning that it is not possible to ban the development of LAWS because of the interests of major powers and the current development of AI weapons, both consider necessary to make these weapons abide to the existing

international laws including the International Humanitarian Law (including principles like proportionality, distinction, traceability and accountability).

Major work is needed to avoid the risks associated with the misuse, the lack of reliability or failed governance and control of AI weapons. For instance, as explained in previous chapters, the amount of non-biased reliable data needed to train AI models and the lack of experience (research and testing) in using these systems in the real world can lead AI systems to make (or recommend) wrong decisions, like launching attacks when it is unnecessary or attack civilians by mistake (illustrating the need for increased reliability and explainability). The oversight and control of humans over these weapons has to be reinforced. The Chinese side proposed for instance to define different levels of autonomy for these weapons and define human control in function of them.

Fu Ying, chairperson of the Centre for International Security and Strategy at Tsinghua University and former vice minister of foreign affairs of China, also recalls that "the various approaches to AI — like behaviourist reinforcement learning, connectionist deep learning and symbolic expert systems — cannot accurately reflect human cognitive capabilities, such as intuition, emotion, responsibility and value" (Allen & Fu 2020). Those cognitive capabilities (we could add empathy), that cannot be modelled today, can be critical in interpreting the situation and in making the right decisions under stress during the battle. For this, and beyond the need for human control and accountability, it is desirable not to completely remove humans from the loop when using AI weapons.

Some real examples may prove this concern to be valid. For instance, in September 1983, an automated early warning system in the USSR reported that the US had launched five land based nuclear missiles. The Soviet officer in charge should have reacted launching himself a nuclear attack which would have started WWIII. However, he suspected ('gut

feeling') that the data was wrong, he waited to confirm the first hit and, by doing this, the world was saved.

Some level of restrain is necessary until these weapons are fully understood. As noted by Fu Ying, "China and the U.S. are in a good position to carry out coordination and cooperation in this area [but] concerns about AI applications are also shared by other countries, which indicates that the challenges are common to humanity and cannot be solved by any one or two countries alone" (Allen & Fu 2020).

More multistakeholder dialog is required to address the existing risks and an international agreement would be necessary to constraint them. As recalled by Allen and West (2021), despite different worldviews and interests, world leaders have been able to reach agreements to constrain certain behaviours and define the rules of war. [Treaties] "provide greater stability and predictability in international affairs, introduce widely held humanitarian and ethical norms into the conduct of war, and reduce the risks of misunderstandings that might spark unintended conflict or uncontrollable escalation".

In December 2021, China, the first country to do so, submitted a position paper on regulating the military applications of AI to the sixth review conference of the UN Convention on Certain Conventional Weapons. The paper recalls that "countries should embrace a vision of common, comprehensive, cooperative and sustainable global security, seek consensus on regulating the military applications of AI […] in order to prevent serious harm or even disasters to mankind caused by military applications of AI" (PRC 2021).

In terms of ethics, the document reminds that "countries need to uphold the common values of humanity, put people's well-being front and centre, follow the principle of 'AI

for good', and observe national or regional ethical norms in the development, deployment and use of relevant weapon systems". In particular, AI weapons should need to respect International Humanitarian Law and other applicable International Laws.

As reported by the Global Times, China called on countries that work on AI military applications "to act in a prudent and responsible manner, refrain from seeking absolute military advantage, and not use AI as a tool to start a war or pursue hegemony. [They should] adhere to the principles of multilateralism, openness, and inclusiveness" (Liu 2021).

These recent developments show that the time to start working in a treaty is now. "The more deeply AI is embedded into military systems and applications before new norms and agreements are reached, the less willingness there will be to roll back any new capabilities they afford, particularly given how costly such systems are to develop" (Allen & Fu 2020).

## 3.4. Proposals for an AI Treaty

As the international community already achieved with the Geneva Conventions, the Chemical Weapons Convention or the Nuclear Non-Proliferation Treaty, a treaty would be necessary to steer, establish limits and create risk mitigation mechanisms for the civilian and military use of AI.

The lessons learned from the Nuclear Non-Proliferation Treaty (NPT) build-up can be beneficial when architecting a treaty on AI. As highlighted by Duarte (2018), "[the NPT] is but a part, a crucial part, of the continuing search for stability and security that can only be attained by means of generally recognized, collectively elaborated, and legally binding

norms that apply equally to every state." The value of the NPT, beyond the agreement itself and the work on building trust during its negotiation, lies also in the regular international meetings that maintain an open dialog to address the concerns or challenges that may arise at any moment.

Nuclear weapons today are used as a deterrence force between major powers. Autonomous weapons, so far, have a much more limited impact but its exponential development makes that soon AI systems could become an existential risk and the deterrence systems will need to be rethought completely (Husain 2021). The world could then revive the 1960's. Nuclear weapons deter because of the horrible consequences of using them. Tegmark (2017) recalls that some may (wrongly) argue that letting nations develop even more horrifying AI based weapons would create even a stronger and wider deterrence system, and maybe end war forever.

Because of the current state-of-the-art of AI weapons, their disruptive capabilities, the fact that the technology is relatively accessible for every country, and it is software based (it can easily be moved around or hidden), it is unrealistic to expect full global ban of their development. However, it would be possible to build a non-proliferation agreement to avoid the development of the most harmful versions of AI systems. According to Frantz (2018), that treaty could be based on the principle by which "nations agree to share the beneficial uses of artificial intelligence and accept universal safeguards to protect against the misuse of these powerful technologies".

The basis for a treaty on AI has been discussed in different forums. John Allen, President of The Brookings Institution and former US Navy General and Darrell West from the

Centre for Technology Innovation at Brookings, published a very comprehensive list of key principles that a treaty on AI should incorporate (Allen & West 2021):

- Incorporate ethical principles such as human rights, accountability, and civilian protection in AI-based military decisions.
- It is vital that humans make the ultimate decisions on missile launches, drone attacks, and large-scale military actions.
- Adopt a norm of 'not' having AI algorithms within nuclear operational command and control systems (or other systems that could be an existential threat for humanity).
- Protect critical infrastructure from digital attacks or AI-powered cyber-weapons.
- Improve transparency on the safety of AI-based weapons systems. Increase reliability, predictability, and stability in weapons development.
- Develop effective oversight mechanisms to ensure compliance with international agreements.

The need to keep humans in the loop is linked to the warnings described in the previous chapter about the lack of capacity of AI to make ethical judgements or leverage human emotions that would prove vital in armed conflict.

Setting the list of critical infrastructures (power grips, health systems…) that should be saved from cyberattacks is an old demand that has not been fulfilled so far. The difficult attribution of these attacks, and therefore the possibility of states hiding behind individuals or activists, was one of the reasons. However, it is worth to clearly define which critical data and infrastructure need to be protected as their disruption unproportionally impact civilians.

Note that regarding the usage of AI with nuclear weapons, there is no real advantage of using AI unless being able to eliminate any possible response for the adversary if attacked first (Boulanin et al. 2019). This is unlikely. Even if an AI controlled system launches an "intelligent" attack or detects earlier than humans an offensive, the consequences would be the same: massive destruction.

The US National Security Commission on Artificial Intelligence report recommendations also include the demand for increased safety (reliability) and transparency (explainability) for AI weapons and the need to forbid AI in nuclear weapons control systems (NSCAI 2021). The commission added that a treaty should also incorporate:

- Venues to discuss AI's impact on crisis stability with competitors.

That is, official and open channels to share concerns and challenges posed by the technology, including the risks of misperceptions.

The Future of Defense Task Force Report (HASC 2020), recalls that any treaty on AI should be:

- Amendable to take into consideration the advancements of the technology (including new threats not considered today).

Finally, a fundamental measure of arms control agreements, critical for building-confidence among all stakeholders and to ensure predictability on weapons development,

is effective and comprehensive verification arrangements (Aboul-Enein 2017). The challenge of AI is that while nuclear weapons require heavy industries, large warehouses and specific materials and technology, an AI-cyberattack can launched from any office, AI weapons just need extra computing power and might be quite similar to conventional ones and AI control systems, from the outside, would be almost undistinguishable from the existing ones.

The ban on biological and chemical weapons is also hard to verify and enforce. But because of the international convention banning them, we have seen few uses of these weapons because of the strong reject and stigmatization that the nation that would use them would suffer. The enforcement challenges should not prevent the development of an AI treaty.

The control and verification of the agreement is the basis for any enforcement measure against those that do not comply to it. It might require the creation of a specialized agency mirroring the International Atomic Energy Agency (IAEA) for instance. Any reinforcement mechanism of the treaty could be addressed through that agency with the support of the UN Security Council.

The current level of international confrontation and the rise of the US-China new bipolarity could complicate reaching agreements on this important topic. But this should not serve as an excuse. After all, at the height of the Cold War, the world reached agreements and treaties on nuclear weapons that proved beneficial for all.

# CONCLUSIONS

> *"AI is humanity's new frontier. Once this boundary is crossed, AI will lead to a new form of human civilization. The guiding principle of AI is not to become autonomous or replace human intelligence. But we must ensure that it is developed through a humanist approach, based on values and human rights. We are faced with a crucial question: what kind of society do we want for tomorrow?"*
>
> Audrey Azoulay, Director-General of the United Nations Educational, Scientific and Cultural Organization, UNESCO (Azoulay 2018).

AI-enabled systems, technology and applications create huge economic, scientific, and social development opportunities but also pose new risks to the society. The capacity of AI to manipulate, socially score and track individuals, and consequently erode fairness, freedoms, justice, and equality, can endanger the healthy functioning of democracies, and strengthen authoritarian regimes.

As the risks, constraints and weaknesses of these technologies are better known, a wide variety of stakeholders have called for the need to regulate AI and establish codes of ethics to be respected during the entire lifecycle of AI systems.

The ethics of AI is a hot topic. In last six months, we have seen as China, the NATO or the UNESCO have all defined AI codes of ethics that add to the work already achieved by the OECD and the EU, just few months before. The EU is leading this normative push by having proposed the first legally binding AI Act. The UNESCO Recommendation, on the other hand, is the first globally accepted recommendation (applicable to civilian applications).

AI has become a moral actor in the social debate moving from being just a technology enabler to become an extended arm of an ideology, a beliefs paladin. The Chinese internet regulator reminded that the algorithms need to protect the communist ideology. The US National Security Commission urges the country to heavily invest in AI to lead the economic (and military) race but also to be able to instil on it 'American' values of freedom, human rights, and democracy. Those values are put forward by the EU Act as well while the UNESCO Recommendation, likely because of its global scope and cultural and scientific skew, only once recommends to 'encourage' democracy.

Although countries around the world have some (sometimes cynical) latitude in interpreting the Human Rights and what Privacy means, all accept them as part of the ethical code that AI should embrace.

The AI ethics requirements have moved from being something that just a few envisioned as necessary to a globally accepted absolute need. The international community cannot accept that AI systems develop beyond direct human oversight, that humans are not in control of it, that AI become unexplainable or unsecure or whereby nobody remain accountable of their acts.

The ethical debate goes beyond a pure moral discussion. The defined development and governance frameworks provide guidance to the industry to build upon, create more interoperable systems and harmonize technical requirements (like reliability and explainability).

The military applications of AI can heavily undermine the current balance of power, challenge the current security order based on conventional military equipment, enable 'low cost' high-damage weapons and armies (accessible to more countries but also terrorist groups), and redefine the security dilemma both in the real and in the cyber-space. The AI 'hyperwar', a light speed war with autonomous weapons which implications no human can really grasp, can be of terrifying consequences for humanity.

The evolution of AI is hard to predict and the dangers that humanity will face or the rules of engagement in the next generation conflicts will be very different from the ones we know today. The currently established deterrence systems might need to be rethought.

The leading nations in AI are well aware of the above. They understand the dangers of developing lethal technology that potentially needs no human to make decisions and no human able to limit its non-intended impact (like harming civilians). A technology that lacks today of the huge amounts of data that are needed to make it reliable in all the theatres of operations. A self-learning technology that we might be unable to understand fully.

Although different think tanks have also called for it previously, China has been the first country to officially request, just few months ago, the creation of an international treaty on regulating AI military applications. While we should never rejoice of having to agree on new rules of war provoked by the human creativity in finding new ways to kill, the push towards the creation of a global treaty is a positive note.

A treaty would bring predictability on this AI arms race and avenues for dialog between adversaries. The NPT was developed at a moment where the humanity faced an existential risk because of the surge of nuclear weapons. The world might be living a similar situation with this new technology.

A unanimous consensus of the international community (at least as voiced by the West and China) is that AI military applications have to abide to the International Humanitarian Law and International Law in general, including the respect of the Human Rights. This should be the starting point for any AI treaty.

This a good opportunity also to address the cyber-risks that have been increasingly growing in complexity and potential harm. With AI, cyberattacks can be more devastating. Being able to agree on the critical infrastructure that needs to be protected and the enforcement mechanisms that should apply in the cyberspace will be valuable.

An AI Treaty should leverage, as much as possible, the ethics code agreed within the framework of the UNESCO. While principles of 'do no harm' may unfortunately not be applicable in military applications, others like explainability, transparency, accountability, human oversight, global governance, etc. are common requirements to the civilian and the military domains. It is important to keep the alignment around common beliefs, development frameworks and ethical expectations in both domains.

On the military side, some principles may need further attention than in the civilian world, like the ones relating to reliability, security, human control, and traceability which are all key into protecting civilians in armed conflicts and respecting International Humanitarian

Law. Additionally, specific cautions or total bans should be agreed on the possibility of using AI algorithms in the control systems of weapons of massive destruction like the nuclear ones.

The creation of a treaty on the usage of AI will encounter numerous difficulties in the making. The current level of global confrontation, the important and disruptive economic potential linked to AI, the new hegemonic duality between US and China and the uncertainties on how this technology may evolve, are some of them. It is also not evident to think on effective mechanisms that could verify that every country complies with the agreed principles. The world demonstrated its capacity to reach agreements at the height of the Cold War. When faced to these new threats, it is worth to try again.

Back to the civilian domain, the next natural step has to be to 'strongly' encourage governments to move from AI ethical recommendations to binding laws that warrant their application. As above, verification and reinforcement of the agreed principles will also need the creation of supervisory boards. The journey from soft to hard laws has already started in the case of the EU. More countries should follow, and the international momentum created by the current agreements around AI ethics should not die.

UN leadership is needed to coordinate the legal initiatives across the globe and to make sure that developing countries do not fall behind. Indeed, any competitive advantage gained by countries that willingly fail to respect the ethical principles should be avoided.

Last but not least, it is important that citizens become more and more versed into AI-led technologies, how algorithms work, and the risks and benefits associated. A

knowledgeable society will be better prepared to strongly defend its freedoms and rights and confront the changes that will come (including the rebalancing of the workforce). Governments should not disregard this essential need.

To summarize, a reliable, trusted, and ethical AI technology, respectful of human dignity, will be a fantastic enabler for a better future, more sustainable, equitable, fair, and less prone to conflict.

Some scientists like Ray Kurzweil forecast that in a near future a technology singularity will bring life to an artificial super-intelligence. Let's hope that this superintelligence when confronted to an existential armed conflict reaches the same conclusion as in the 1983 classic American film "WarGames". In that movie, Matthew Broderick plays the role of a young hacker that accidentally gets access to the US military supercomputer called WOPR (War Operation Plan Response) programmed to simulate, predict, and execute nuclear war against the Soviet Union. Does it sound familiar?

Due to security failures (reliability!), the WOPR thinks that the US is under a real nuclear attack. Before initiating the counterattack, launching the US nuclear arsenal, and actually starting WWIII, the computer simulates (unsupervised learning) all the possible nuclear war scenarios finding that all of them lead to a mutual assured destruction (MAD).

The AI stops the missile launching sequence and reports its conclusion: "This is a strange game. The only winning move is not to play."

# REFERENCES


- Aboul-Enein, S. 2017, "Toward a Non-Nuclear World: The NPT Regime - Nuclear Disarmament and the Challenge of a WMDFZ in the Middle East". In: *International Journal of Nuclear Security*, vol. 3, no. 1, 2017

- Alan Turing Institute, (ATI) 2021, “Artificial Intelligence, Human Rights, Democracy and the Rule of Law” https://edoc.coe.int/en/artificial-intelligence/10206-artificial-intelligence-human-rights-democracy-and-the-rule-of-law-a-primer.html

- Allen, J. R., 2018, “Interview with General John R. Allen, USMC (Ret.).” PRISM, vol. 7, no. 4, Institute for National Strategic Security, National Defense University, 2018, pp. 148–54, https://www.jstor.org/stable/26542713.

- Allen, J.R., Fu, Y., 2020, "Together, The U.S. And China Can Reduce The Risks From AI", *Noema, The Berggruen Institute*, 17 December 2020, https://www.noemamag.com/together-the-u-s-and-china-can-reduce-the-risks-from-ai/

- Allen, J.R., West, D., 2021, "It is time to negotiate global treaties on artificial intelligence", Brookings, 24 March 2021, https://www.brookings.edu/blog/techtank/2021/03/24/it-is-time-to-negotiate-global-treaties-on-artificial-intelligence/

- Arkin, R.C., 2015, "Warfighting Robots Could Reduce Civilian Casualties, So Calling for a Ban Now Is Premature", *IEEE Spectrum*, 4 August 2015,

https://spectrum.ieee.org/autonomous-robotic-weapons-could-reduce-civilian-casualties

- Arkin, R.C., Kaelbling, L., Russell, S., Sadigh, D., Scharre, P., Selman, B., Walsh, T., 2019, "A Path Towards Reasonable Autonomous Weapons Regulation", *IEEE Spectrum*, 21 October 2019, https://spectrum.ieee.org/a-path-towards-reasonable-autonomous-weapons-regulation

- Arnett, E. H., 1992. "Welcome to Hyperwar". *Bulletin of the Atomic Scientists*, 48(7), 14–21. https://doi.org/10.1080/00963402.1992.11460097

- Azoulay, A. 2018, “Towards an Ethics of Artificial Intelligence”, UN Chronicle, December 2018, Nos. 3 & 4 Vol. LV https://www.un.org/en/chronicle/article/towards-ethics-artificial-intelligence

- Bendett, S., 2017, “Red Robots Rising: Behind the Rapid Development of Russian Unmanned Military Systems”, *The Strategy Bridge*, 12 December 2017, https://thestrategybridge.org/the-bridge/2017/12/12/red-robots-rising-behind-the-rapid-development-of-russian-unmanned-military-systems

- Boulanin, V. (editor) et al., 2019 “The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk” Volume I, Euro-Atlantic Perspectives, Stockholm International Peace Research Institute (SIPRI)

- Boulanin, V., Ruun, L. and Goussac, N., 2021, *Autonomous Weapon Systems and International Humanitarian Law*, Stockholm International Peace Research Institute (SIPRI)

- Buchanan, B., 2020, “The AI Triad and What It Means for National Security Strategy”, Center for Security and Emerging Technology (CSET) at Georgetown’s Walsh School of Foreign Service, August 2020
- CAHAI, Ad-hoc Committee on Artificial Intelligence, 2020, “Towards regulation of AI systems. Global perspectives on the development of a legal framework on Artificial Intelligence systems.” https://edoc.coe.int/en/artificial-intelligence/9656-towards-regulation-of-ai-systems.html
- CCW Convention 2019, Group of Governmental Experts (GGE) on Emerging Technologies in the Area of LAWS, 2019, “Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems”, CCW/GGE.1/2019/3, 25 Sep. 2019, Annex IV
- Clark, C., 2017, "Our Artificial Intelligence ‘Sputnik Moment’ Is Now: Eric Schmidt & Bob Work", *Breaking Defense*, 1 November 2017, https://breakingdefense.com/2017/11/our-artificial-intelligence-sputnik-moment-is-now-eric-schmidt-bob-work/
- Constantinou, C. M., 2013, "Between Statecraft and Humanism: Diplomacy and Its Forms of Knowledge". *International Studies Review*, Vol. 15, No. 2
- Department of Defense US (DoD), 2019, "Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense", https://media.defense.gov/2019/Oct/31/2002204459/-1/-1/0/DIB_AI_PRINCIPLES_SUPPORTING_DOCUMENT.PDF

- Detsch, J., 2021, "The U.S. Army Goes to School on Nagorno-Karabakh Conflict", *Foreign Policy*, 30 March 2021, https://foreignpolicy.com/2021/03/30/army-pentagon-nagorno-karabakh-drones/

- Duarte, S., 2018, "Unmet Promise: The Challenges Awaiting the 2020 NPT Review Conference". In: *Arms Control Today*; Vol. 48 Issue 9

- EC, European Commission, Directorate-General for Communications Networks, Content and Technology, 2021, "Proposal for a Regulation of the European Parliament and of the Council, laying down harmonised rules on Artificial Intelligence (AI Act) and amending certain Union legislative Acts", 21 April 2021, COM/2021/206 final, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

- Etzioni, A. 2018, "Pros and Cons of Autonomous Weapons Systems" in: *Happiness is the Wrong Metric.* Library of Public Policy and Public Administration, vol 11. Springer, Cham. https://doi.org/10.1007/978-3-319-69623-2_16

- Etzioni, O., Decario, N., 2019, "We have the basis for an international AI treaty", *The Hill*, 17 July 2019, https://thehill.com/opinion/technology/452809-we-have-the-basis-for-an-international-ai-treaty

- Frantz, D., 2018, "We've unleashed AI. Now we need a treaty to control it", Los Angeles Times; 16 July 2018, https://www.latimes.com/opinion/op-ed/la-oe-frantz-artificial-intelligence-treaty-20180716-story.html

- Future of Life Institute (FoL), 2017, “Asilomar AI Principles”, https://futureoflife.org/2017/08/11/ai-principles/
- Gaumond, E., 2021, “Artificial Intelligence Act: What Is the European Approach for AI?”, *Lawfare*, 4 June 2021, https://www.lawfareblog.com/artificial-intelligence-act-what-european-approach-ai
- Global Fire Power (GFP), 2022, "Defense Spending by Country (2022)", consulted on 1 March 2022, https://www.globalfirepower.com/defense-spending-budget.php
- Guterl, F., 2020, "As China Leads Quantum Computing Race, U.S. Spies Plan for a World with Fewer Secrets", Newsweek, 14 December 2020, https://www.newsweek.com/2020/12/25/china-leads-quantum-computing-race-us-spies-plan-world-fewer-secrets-1554439.html
- Hernandez, J., 2021, "A Military Drone With A Mind Of Its Own Was Used In Combat, U.N. Says", *NPR*, 1 June 2021, https://www.npr.org/2021/06/01/1002196245/a-u-n-report-suggests-libya-saw-the-first-battlefield-killing-by-an-autonomous-d?t=1646746242705
- Horowitz, M.C., 2018, "Artificial Intelligence, International Competition, and the Balance of Power", *Texas National Security Review:* Volume 1, Issue 3, https://doi.org/10.15781/T2639KP49, https://tnsr.org/2018/05/artificial-intelligence-international-competition-and-the-balance-of-power
- House Armed Services Committee (HASC), 2020 "Future of Defense Task Force Report 2020",

https://armedservices.house.gov/_cache/files/2/6/26129500-d208-47ba-a9f7-25a8f82828b0/6D5C75605DE8DDF0013712923B4388D7.future-of-defense-task-force-report.pdf

- Hughes, R., 2010, “A treaty for cyberspace”. *International Affairs*, 86(2). 2010 p.523-529. ISSN 00205850

- Human Rights Watch (HRW), 2020, "Stopping Killer Robots: Country Positions on Banning Fully Autonomous Weapons and Retaining Human Control", 10 August 2020, https://www.hrw.org/sites/default/files/media_2021/04/arms0820_web_1.pdf

- Hunt, E., 2016, "Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter", *The Guardian*, 24 March 2016, https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter?CMP=twt_a-technology_b-gdntech

- Husain, A., 2021, “AI Is Shaping the Future of War.” PRISM, vol. 9, no. 3, Institute for National Strategic Security, National Defense University, 2021, pp. 50–61, https://www.jstor.org/stable/48640745.

- International Research Center (IRC) for AI Ethics and Governance, 2019, "Beijing Artificial Intelligence Principles", https://ai-ethics-and-governance.institute/beijing-artificial-intelligence-principles/

- International Research Center (IRC) for AI Ethics and Governance, 2021, "The Ethical Norms for the New Generation Artificial Intelligence, China", https://ai-

ethics-and-governance.institute/2021/09/27/the-ethical-norms-for-the-new-generation-artificial-intelligence-china/

- International Telecommunication Union (ITU), 2021, "United Nations Activities on Artificial Intelligence (AI) 2021", https://www.itu.int/hub/publication/s-gen-unact-2021/

- Jasper, M., 2021, “U.S. Unprepared for AI Competition with China, Commission Finds”, *Nextgov*, 1 March 2021, https://www.nextgov.com/emerging-tech/2021/03/us-unprepared-ai-competition-china-commission-finds/172377/

- Jones, B., Feltman, J., and Moreland, W. 2019. “Competitive Multilateralism: Adapting Institutions to meet the New Geopolitical Environment”. Foreign Policy, Brookings Institute. 20 September

- Kharpal, A., 2021, "China wants to be a $150 billion world leader in AI in less than 15 years", *CNBC*, 21 July 2021, https://www.cnbc.com/2017/07/21/china-ai-world-leader-by-2030.html

- Kirkpatrick, D.D., Frenkel, S., 2017, "Hacking in Qatar Highlights a Shift Toward Espionage-for-Hire", *New York Times*, 8 June 2017, https://www.nytimes.com/2017/06/08/world/middleeast/qatar-cyberattack-espionage-for-hire.html?smid=em-share

- Krishna, A., 2020, “IBM CEO’s Letter to Congress on Racial Justice Reform”, 8 June 2020, https://www.ibm.com/blogs/policy/facial-recognition-sunset-racial-justice-reforms/

- Lant, K., 2017, "China Aims to be a Global Frontrunner in AI by 2030", *Futurism*, 24 July 2017, https://futurism.com/china-aims-to-be-a-global-frontrunner-in-ai-by-2030

- Liu, X. 2021, "China urges regulating military use of AI, first time in UN history, showing global responsibility", *Global Times*, 14 December 2021, https://www.globaltimes.cn/page/202112/1241470.shtml

- Lucas, G.R., Jr. 2013. "Engineering, ethics & industry: The moral challenges of lethal autonomy" in *Killing by remote control: The ethics of an unmanned military*, ed. B. Strawser, 211–228. Oxford: Oxford University Press.

- Mahbubani, K., 2013, "Multilateral Diplomacy". In: Cooper, A.F. et al. *The Oxford Handbook of Modern Diplomacy*, Oxford University Press, Oxford

- McCarthy, J., Minsky, M.L., Rochester, N., Shannon, C.E., 1955. "A proposal for the Dartmouth Summer Research Project on Artificial Intelligence", *AI Magazine*, 2016, vol. 27, no. 4, pp.12-14.

- Meserole, C., 2018, "Artificial intelligence and the security dilemma", *Brookings*, 6 November 2018, https://www.brookings.edu/blog/order-from-chaos/2018/11/06/artificial-intelligence-and-the-security-dilemma/

- National Security Commission on Artificial Intelligence (NSCAI), USA, 2021, "Final Report", October 2021, https://www.nscai.gov/

- North Atlantic Treaty Organization (NATO), 2021, "Summary of the NATO Artificial Intelligence Strategy", https://www.nato.int/cps/en/natohq/official_texts_187617.htm

- Nye, J. S.; 2017, "Deterrence and Dissuasion in Cyberspace". *International Security* 2017; 41 (3): 44–71. doi: https://doi.org/10.1162/ISEC_a_00266
- Organization for Economic Co-operation and Development (OECD), 2019, "Recommendation of the Council on Artificial Intelligence", OECD/LEGAL/0449, Adopted 22 May 2019, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449
- Pauwels, E. 2018, "How Can Multilateralism Survive the Era of Artificial Intelligence?", UN Chronicle, December 2018, Nos. 3 & 4 Vol. LV, https://www.un.org/en/chronicle/article/how-can-multilateralism-survive-era-artificial-intelligence
- PRC, Permanent Mission of People's Republic of China to the UN, Office at Geneva, 2021, "Position Paper of the People's Republic of China on Regulating Military Applications of Artificial Intelligence (AI)"; 13 December 2021, http://www.china-un.ch/eng/dbtyw/cjjk/202112/t20211213_10467517.htm
- Riordan, S. 2018. "The Geopolitics of Cyberspace: a Diplomatic Perspective", *Brill Research Perspectives in Diplomacy and Foreign Policy*, 3(3), 1-84. doi: https://doi.org/10.1163/24056006-12340011
- Ruggie, J. G. 2015. "Life in the Global Public Domain: Response to Commentaries on the UN Guiding Principles and the Proposed Treaty on Business and Human Rights.". At SSRN: doi:10.2139/ssrn.2554726.
- Schmitt, M. 1998, "Bellum Americanum", in *The law of armed conflict into the next millennium*, Newport, Naval War College

- Shen, X., 2021, "Chinese AI gets ethical guidelines for the first time, aligning with Beijing's goal of reining in Big Tech", *South China Morning Post*, 3 October 2021, https://www.scmp.com/tech/big-tech/article/3150789/chinese-ai-gets-ethical-guidelines-first-time-aligning-beijings-goal

- Shen, X., Qu, T., 2021 China draws up plan to bring algorithms under state control in sign of tightened censorship", *South China Morning Post*, 9 September 2021, https://www.scmp.com/tech/policy/article/3150608/china-draws-plan-bring-algorithms-under-state-control-sign-tightened?module=inline&pgtype=article

- Snow, S., 2020, "The Corps is axing all of its tank battalions and cutting grunt units", *Marine Corps Times*, 23 March 2020, https://www.marinecorpstimes.com/news/your-marine-corps/2020/03/23/the-corps-is-axing-all-of-its-tank-battalions-and-cutting-grunt-units/

- Tegmark, M., 2017, "LIFE 3.0. Being Human in the Age of Artificial Intelligence", Vintage Books, New York

- United Nations Educational, Scientific and Cultural Organization (UNESCO), 2019 "Preliminary study on the technical and legal aspects relating to the desirability of a standard-setting instrument on the ethics of artificial intelligence" Executive Board 206 EX/42 https://unesdoc.unesco.org/ark:/48223/pf0000367422

- United Nations Educational, Scientific and Cultural Organization (UNESCO), 2021, "Recommendation on the ethics of artificial intelligence" SHS/BIO/REC-AIETHICS/2021 https://unesdoc.unesco.org/ark:/48223/pf0000380455

- Valeriano, B., Maness, R.C., 2018. "International Relations Theory and Cyber Security: Threats, Conflicts, and Ethics in an Emergent Domain" in *The Oxford Handbook of International Political Theory*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198746928.013.19

- Villani, C., Schoenauer, M., Bonnet, Y., Berthet, C., Cornut, A.C., Levin, F., Rondepierre, B., 2018, "For a Meaningful Artificial Intelligence", *Mission Villani sur l'intelligence artificielle*, https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf

- Vincent, J. 2017 "Putin Says the Nation That Leads in AI 'Will Be the Ruler of the World,'" *The Verge*, Sept. 4, 2017, https://www.theverge.com/2017/9/4/16251226/russia-ai-putin-rule-the-world

- Wiseman, G. 2015, "Diplomatic practices at the United Nations". Cooperation and Conflict, Vol. 50(3), 2015. ISSN 00108367

- World Economic Forum (WEF), 2019, "AI Government Procurement Guidelines", https://www.weforum.org/whitepapers/ai-government-procurement-guidelines