International Journal of Humanities and Social Science Invention (IJHSSI)
ISSN (Online): 2319 – 7722, ISSN (Print): 2319 – 7714
www.ijhssi.org ||Volume 9 Issue 6 Ser. V || June 2020 || PP 30-42

The very useful proposal of the European Union for Shaping Artificial Intelligence

Celso Vargas Elizondo
Instituto Tecnológico de costa Rica

Summary: *In this paper the European proposal for shaping Artificial Intelligence(AI) is reviewed. It is based on three relevant documents published by the European Union, one of them written by a specialized team. It provides very useful suggestions on how to discuss and propose frameworks for boosting and directing the development of AI in our countries. The paper divides into three main sections. In section one, a brief introduction to AI is presented. The second section introduces the European plan for boosting AI in Europe. And the third section introduces the ethical guidelines proposed by European Union for shaping AI. The most extend part of the paper is devoted to these ethical guidelines.*

Key words: *Artificial Intelligence, Ethical Guidelines, European Union, Applied Ethics*

Date of Submission: 18-06-2020 Date of Acceptance: 04-07-2020

The European Union had published, recently, several important documents on Artificial Intelligence (AI) for the Eurozone. Diverse and relevant areas of application of AI are considered in these publications as well as a framework for AI. Among the documents, three of them are relevant to our review of the role that AI will play in Europe: *Communication Coordinated Plan on Artificial Intelligence (COM(2018) 795 final)*released in February 2020; *Ethics Guidelines for Trustworthy AI (*the second draft published in 2019, the final version will be available during this year). And *On Artificial Intelligence - A European approach to excellence and trust*(2020). The first and the third overlap in several issues. The Coordinated Plan presented in *On Artificial Intelligence*and in*Communication Coordinated Plan*is the umbrella under which other documents can be framed. This coordinated plan is based in three pillars: "increasing public and private investments in AI, preparing for socio-economic changes, and ensuring an appropriate ethical and legal framework. To ensure its success, coordination at The European level is essential" (EU, 2020a:2). Ideas and proposals found in these documents are results of the requests and decisions made previously in the EU, some of them go back several years ago. A more comprehensive analysis of The European Union position on this issue should take into account these antecedents, but it is not our purpose in this paper.

But before presenting The European perspective, it is important to talk a little about AI aimed at understanding the relevant role that this issue is taking in Europe.

I. Artificial Intelligence

Several AI algorithms are routinely used, for example, in classifying an email as spam or not, in search engine results pages (SERP), voice recognition assistant systems such as Alexa and Siri, voice to text conversion, natural language modelling, face recognition, speech recognition, technical costumer assistants, and in many available apps. More complex algorithms include, for example, autonomous driving, specialized robots for industrial processes, and assistant robots in medicine. However, the era of AI is just starting, many theoretical problems should be solved in the near future, but the potentiality of this research and development area is enormous. AI systems work better with big data, so as data increases, their performance will be better. On the other hand, one important achievement of AI is that these systems do very well in structured, non-structured probabilistic and non-probabilistic environments.

However, there is no consensus yet on what is AI. EU proposal follow Russell and Norvig (2009) definition of AI:

"Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals.

AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications)" (AI-HLG, 2019b, 1).

There exist two main research paradigms for yielding intelligent behavior artificially. The first one is called symbolic or one-level paradigm that attempts to capture and programme intelligent behaviors in machines by means of declarative formalisms. The most representatives of this paradigm are the logical based formalisms. Several logical formalisms are used, including propositional and first order logical calculus, and different higher order logical formalisms, both modal and non-modal, to capture different properties of knowledge representation and inference, and to model different intelligent structures. Non-monotonic logics are the most appropriate for representing and modelling domains such as reasoning and common sense. Finally, it plays an important role as programming language (see Thomason, 2020). However, there is another important use of logic in AI and in computer science in general: It provides the base for evaluating the correctness of different implementations independently of the programming language used.

The second paradigm is called sub-symbolic or multi-layered paradigm. For this paradigm intelligent behavior is an emergent property. It emerges from the micro programming or neuro-computation of the system. Neural networks and deep learning belong to this category. This paradigm receives an important impulse during the 80´s and 90´s of the XX century, with the influential publication of Rumelhart and McCleland in 1986; two volumes dedicated to Parallel Distributed Processing (PDP). This approach was called "connectionism" because it pretends to simulate computationally the way in which our brain processes information, represent knowledge, and makes decisions. However, connectionism is currently more closely connected with cognitive science, also an important field of research in AI, computer science and psychology, among other fields (see Buckner and Garson (2019) for a recent account of this research field).

These paradigms can be called "pure paradigms", in the sense that the ambitious of them are to arrive to an artificial general intelligence (AGI) as will be discussed below. This ambitious could also be shared by the different approaches within each paradigm. And of course, some researchers are more pragmatic and don´t follow a "pure" paradigm (see Bringsjord and Govindarajulu, 2020 section 3.4 "AI Beyond the Clash of Paradigms").

However, as diverse as these approaches and paradigms could be, there is a convergent point: the **intelligent agent**. An intelligent agent is one that exhibits a behavior that is intelligent when evaluated by a human being. Of course, a "human being" is the model used to attribute intelligence to a computer programme. For a computer programme to be intelligent, according to human model, has to have the following attributes: language and visual performance, memory, abstract and practical reasoning and reacting to the environmental stimuli, among others. All these human intelligent capabilities are of permanent research in AI. In some cases, these machines clearly surpass human beings: their ability to process large amount of information, memory storage and retrieval capacity and in abstract reasoning, such as chess.

This concept of intelligent agent is very important in AI. As mentioned above this is called artificial general intelligence (AGI). It is one of the current research area. However, this idealistic or universal intelligent agent is difficult to implement in the near, even in the long term. Because of this, more modest AI systems were proposed after the 1990. Then, a less strict concept of "artificial agent" was also proposed. One of the most influential book on AIis Russell and Norvig (2009) introduces the intelligent agent in the following way:

"The main unifying theme is the idea of an intelligent agent. We define AI as the study of agents that receive percepts from the environment and perform actions. Each such agent implements a function that maps percept sequences to actions, and we cover different ways to represent these functions… (Russell &Norvig 2009, vii)

More than one of these functions are required in some specific application. For example, autonomous driving distinct functions are needed, among them: some group for immediate external percepts to sensor and act accordingly; some other functions are needed to evaluate the current state of the machine and its

components, and others to take information from the could (as GPS, Galileo or Glonass and its interoperability) to drive to the desirable destination.

This more restricted definition of AI has been very successful. As pointed out by Acemoglu andRestrepo,

"The big breakthroughs and the renewed excitement in AI are coming from advances in hardware and algorithms that enable the processing and analysis of vast amounts of unstructured data (for example, speech data that cannot be represented in theusualstructuredways, suchasinsimple, Excel-likedatabases). Centraltothisrenaissance of AI have been methods of machine learning (which are the statistical techniques that enable computers and algorithms to learn, predict and perform tasks from large amounts of data without being explicitly programmed) and what is called deep learning(algorithms that use multi-layered programs, such as neural nets, for improved machine learning, statistical inference and optimization)" (Acemoglu and Restrepo, 2019: 3).

So, flexibility and different degrees of integration are part of the strategy followed in this pragmatic momentum in approaching AI.

This very brief reference to this relevant field of AI aims at providing a context to better understand EU proposal for shaping AI within a robust ethical framework.

II. The European Union approach to promote AI

Coordinated Plan on Artificial Intelligence (CP). Given thatmany references are made in it to the other two documents, I will make a very brief reference to this plan. The plan starts referring to more relevant strengths that EU has and makes it realistic to propose an ambitious project on AI. Among them: Very good research centres and universities, good trained human resources, many startups, industries, including "robotics and competitive manufacturing and services sectors, from automotive to healthcare, energy, financial services and agriculture. Europe has developed a strong computing infrastructure (e.g. high-performance computers), essential to the functioning of AI" (CP, 4). However, EU lacked of an agreed plan to promote and regulate AI in all the members of the EU.

"A common European approach to AI is necessary to reach sufficient scale and avoid the fragmentation of the single market. The introduction of national initiatives risks to endanger legal certainty, to weaken citizens' trust and to prevent the emergence of a dynamic European industry." (CP, 3).

So, CP considers the following four essential components to be reached in this boosting process of enhancing EU position in AI: 1) Capitalising on strengths in industrial and professional markets; 2) Seizing the opportunities ahead: the next data wave, 3) An ecosystem of excellence, 4) An ecosystem of trust: regulatory framework for AI. I will refer only to the first three components, because the other is better presented in the documents *On Artificial Intelligence* and *Ethics Guidelines for Trustworthy AI.*

Related to the first component, "Capitalisingon strengths in industrial and professional markets" two aspects should be mentioned: First, in this plan it is not included any prevision for the application of AI in defense, development of new arms, and massive destructive devices/weapons. There is no reference to the use of AI on this matter. Instead it is explicitly excluded from the The European priority areas in which AI will be promoted. In the *Guidelines for Trustworthy AI,*this kind of warlike applications are used to exemplify the "critical concerns raised by AI". For example, referring to lethal autonomous weapon systems (LAWS), in the *Guidelines*, we read:

"Currently, an unknown number of countries and industries are researching and developing lethal autonomous weapon systems, ranging from missiles capable of selective targeting to learning machines with cognitive skills to decide whom, when and where to fight without human intervention. This raises fundamental ethical concerns, such as the fact that it could lead to an uncontrollable arms race on a historically unprecedented level, and create military contexts in which human control is almost entirely relinquished and the risks of malfunction are not addressed. The European Parliament has called for the urgent development of a common, legally binding position addressing ethical and legal questions of human control, oversight,

accountability and implementation of international human rights law, international humanitarian law and military strategies. Recalling the European Union's aim to promote peace as enshrined in Article 3 of the Treaty of the European Union, we stand with, and look to support, the Parliament's resolution of 12 September 2018 and all related efforts on LAWS." (AI-HLG, 2019: 34).

So, the emphasis of this plan orbits around the centrality of humans and the protection of the environment.

Secondly, the scope of application of the Coordinated Plan divides into three main categories:

a) Citizens in areas such as "health care, fewer breakdowns of household machinery, safer and cleaner transport systems, better public services"
b) Business in those areas in which Europe is now strong, such as "machinery, transport, cybersecurity, farming, the green and circular economy, healthcare and high-value-added sectors like fashion and tourism"
c) Public services, such as, "reducing the costs of providing services (transport, education, energy and waste management), by improving the sustainability of products and by equipping law enforcement authorities with appropriate tools to ensure the security of citizens, with proper safeguards to respect their rights and freedoms" (CP 2020a, 3).

To invest in the development of AI in these three categories is particularly relevant in the sense that this could represent a turning point in the approach to AI. Acemoglu and Restrepo (2019) analyse the current tendency in the development of AI worldwide, characterized by an excessive emphasis on automation of industrial and services processes. According to them,

"The standard approach, both in popular discussions and academic writings, presumes that any advance that increases productivity (value added per worker) also tends to raise the demand for labor, and thus employment and wages. Of course, technological progress might benefit workers with different skills unequally and productivity improvements in one sector may lead to job loss in that sector. But even when there are sectoral job losses, the standard narrative goes, other sectors will expand and contribute to overall employment and wage growth." (Acemoglu and Restrepo, 2019: 3).

They questioned this "automatic connection between automation and productivity" and suggest, on contrary, that available data shows that the introduction to automation in the production and service processes reduce the number of jobs needed for keeping the same productivity. Acemoglu and Restrepo don´t quote directly proposals such as that of World Bank, but it is clear that this organization endorses such positions. According the last report (2019) called *The Changing Nature of Work*,

"Creating formal jobs is the first-best policy, consistent with the International Labour Organization's decent work agenda, to seize the benefits of technological change. In many developing countries, most workers remain in low-productivity employment, often in the informal sector with little access to technology." (World Bank, 2019: 4).

Instead, what is important is to invest in those new areas in which AI can improve the quality of life of people and, at the same time, permit the creation of new jobs. What Acemoglu and Restrepo recommend is to separate both issues. It is important to identify new niches for automation because this is one of strengthens of AI, but at the same time to look for the creation of new jobs that benefits of these new niches. Particularly relevant here are those AI developments that have the potential to allow that innovation arises in diverse forms within the country or region. But to accomplish this, an active role of government and their relevant institutions is very relevant, not only in the identification of these niches but also in investment to permit a better harmonization in the penetration of these technologies.

The second component, "Seizing the opportunities ahead: the next data wave", begins by recognizing the weak position of EU in "consumer applications and on online platforms" but at the same time the opportunities that this area offers to EU to change the balance between the centralized used of information in the cloud, and that "in smart connected objects, such as cars, home appliances or manufacturing robots, and in computing facilities close to the user". Currently the ratio is 80/20. The three categories mentioned above have a potential to further moving the use of massive data closer to end-users. Three main areas are relevant here the change the balance: quantum computing in which EU is ahead; the leading position of EU in machine learning

and deep learning, and third, in symbolic approached to AI. "Combining symbolic reasoning with deep neural networks may help us improve explainability of AI outcomes" (CP, 2020a, 6).

In connection to "the constellation of excellence", the plan, introduces eight different levels of action. A) *Working with member states* includes about "70 joint actions for closer and more efficient cooperation between Member States" in the key areas relevant for achieving the goal; among them, "research, investment, market uptake, skills and talent, data and international cooperation". B) *Focusing the efforts of the research and innovation community* is oriented to create synergies in the Eurozone for boosting an AI of excellence, including a network of centres of research located in strategic places in Europe, and a "centre of research, innovation and expertise" that coordinates all the efforts in the whole region. C) *Skills* include an updated Digital Education Action Plan to improve the use of data and AI related technologies in all the levels of Europe educational system. One important goal in this plan is attract talents in this area from overseas. D) Focus on SMEs. Small and Medium Enterprises is one important contributor to the economy of the zone. Initiatives like Digital Innovation Hubs and The European Investment Fund are devoted to improve SMEs access and benefits of AI development and deployment. E) Partnership with the private sector. It is included a full involvement of the sector and also to bank funds from the private sector to AI research and development activities. F) Promoting the adoption of AI by the public sector. Here the focus is "in the areas of healthcare and transport where technology is mature for large-scale deployment.". G)Securing access to data and computing infrastructures. This level of action complements other plans already existents in Europe, particularly, The European data strategy and also funding support provisions from the Digital Europe Programme to the priorities defined in this coordinated plan. H) International aspects. As it is indicated in this level of action, Europe has a very long tradition in cooperation with different counterparts worldwide, including G20, UN, OECD, WTO and many others. This cooperation will be enhanced with the implementation of this plan. Another interesting issue concerns, and it is included in the plan, the potentialities of AI to contribute to achieve the "Sustainable Development Goals and advance the 2030 Agenda."

III. An ecosystem of trust: regulatory framework for AI

This fourth area of the CP is developed in the *Ethics Guidelines for Trustworthy AI*. Requirements for the elaboration of these guidelines come from the CP. These requirements are a structural part of the guidelines and are the following:

- "Human agency and oversight,
- Technical robustness and safety,
- Privacy and data governance,
- Transparency,
- Diversity, non-discrimination and fairness,
- Societal and environmental wellbeing, and
- Accountability" (CP,2020a, 10)

In this sense, these guidelines are a structural part of CP. However, the guidelines include a first layer from which the proposed requirements are derived, called, "Foundations of trustworthy AI". This level is principle-oriented and attempts, as we will see, to connect these guidelines to more fundamental values and regulations of the EU.

These guidelines are structured following a three-layer strategy, called, "foundation of trustworthy AI", "Realising Trustworthy AI" and "Assessing Trustworthy AI", respectively. The system has the shape of a cascade, in which, the first layer is the base on which the second layer is based, and the second one, is the base for the third layer. So, it goes from a more abstract perspective to a more concrete one. It is a very consistent approach as we will show here. Additionally,

"Trustworthy AI has three components, which should be met throughout the system's entire life cycle: (1) it should be lawful, complying with all applicable laws and regulations (2) it should be ethical, ensuring adherence to ethical principles and values and (3) it should be robust, both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm" (AI-HLG, 2019: 4).

As observed, there are two dimensions in these guidelines. The first on those ethical and legal issues to make coherent the proposal. The second, on the social and institutional structure and procedures for making transparent, accountable and explicable the decisions making processes.

Finally, it is included in the proposal "examples of opportunities and critical concerns raised by AI" in which the authors discussed both opportunities for AI in EU and those developments that should yield higher concern in EU, because they move away from The European values or civilian applications of AI.

Let us introduce in more detail each component of this interesting proposal.

1) **Foundation of trustworthy AI.**In this component three aspects are discussed: the fundamental principles on which trustworthy AI is founded, the emphasis on asymmetries on the development, deployment and used of AI systems, and unexpected negative impacts of AI in society.

Four are the fundamental principles adopted the AI-HLG: respect for human autonomy, prevention of harm, fairness and explicability. These four concepts are grounded in the foundational regulations of EU, in international treaties approved by EU and in other related regulations. According to it,

"Legal sources include, but are not limited to: EU primary law (the Treaties of the European Union and its Charter of Fundamental Rights), EU secondary law (such as the General Data Protection Regulation, the Product Liability Directive, the Regulation on the Free Flow of Non-Personal Data, anti-discrimination Directives, consumer law and Safety and Health at Work Directives), the UN Human Rights treaties and the Council of Europe conventions (such as the European Convention on Human Rights), and numerous EU Member State laws. Besides horizontally applicable rules, various domain-specific rules exist that apply to particular AI applications (such as for instance the Medical Device Regulation in the healthcare sector)" (AI-HLG, 2019b, 8).

Autonomy, prevention of harm, fairness and explicability impregnate all these regulations, and are the source for new regulations and also for the up-dating of the existent ones.

We have differentiated between those understood as concepts from principles. Generally speaking, a principle is a statement that provides some guidance on the way in which the involved people should behave or the way in which she has to proceed. For example, on respecting autonomy as principle it is formulated in the following way:

"The fundamental rights upon which the EU is founded are directed towards ensuring respect for the freedom and autonomy of human beings. Humans interacting with AI systems must be able to keep full and effective self-determination over themselves, and be able to partake in the democratic process. AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans. Instead, they should be designed to augment, complement and empower human cognitive, social and cultural skills. The allocation of functions between humans and AI systems should follow human-centric design principles and leave meaningful opportunity for human choice. This means securing human oversight over work processes in AI systems. AI systems may also fundamentally change the work sphere. It should support humans in the working environment, and aim for the creation of meaningful work."(AI-HLG, 2019b, 14).

So, respect for human autonomy are related to other fundamental rights and values present in human rights declarations, in almost all constitutions and in other national legal regulations. But at the same time, human autonomy is important in all of the modern ethical approaches. As presented in these guidelines, "respect for human autonomy" is closely related to the following values and rights: respect for human dignity, freedom of the individual, Respect for democracy, justice and the rule of law, the rights of persons at risk of exclusion and Citizens' rights.

The principle of prevention of harm is closely related to the principle of respect human autonomy. It is like the heads and tails in a coin. In general, preventing harm entails to pay attention to those situations and acts that can negatively affect or limit human autonomy and dignity. However, in this proposal the emphasis is on asymmetries and more vulnerable sectors and individuals of society.

"AI systems should neither cause nor exacerbate harm or otherwise adversely affect human beings. This entails the protection of human dignity as well as mental and physical integrity. AI systems and the environments in which they operate must be safe and secure. They must be technically robust and it should be ensured that they are not open to malicious use. Vulnerable persons should receive greater attention and be included in the development, deployment and use of AI systems. Particular attention must also be paid to

situations where AI systems can cause or exacerbate adverse impacts due to asymmetries of power or information, such as between employers and employees, businesses and consumers or governments and citizens. Preventing harm also entails consideration of the natural environment and all living beings." (AI-HLG, 2019b, 14)

While these two principles are oriented to respecting and preventing human autonomy, the other two principles provide guidelines on the way in which the social distribution of costs and benefits are assured, and on the traceability and transparency of the system as a whole.

The principle of fairness is formulated in the following way:

"The development, deployment and use of AI systems must be fair. While we acknowledge that there are many different interpretations of fairness, we believe that fairness has both a substantive and a procedural dimension. The substantive dimension implies a commitment to: ensuring equal and just distribution of both benefits and costs, and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatisation. If unfair biases can be avoided, AI systems could even increase societal fairness. Equal opportunity in terms of access to education, goods, services and technology should also be fostered. Moreover, the use of AI systems should never lead to people being deceived or unjustifiably impaired in their freedom of choice. Additionally, fairness implies that AI practitioners should respect the principle of proportionality between means and ends, and consider carefully how tobalance competing interests and objectives. The procedural dimension of fairness entails the ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them. In order to do so, the entity accountable for the decision must be identifiable, and the decision-making processes should be explicable.(AI-HLG, 2019b, 14-15)

Finally, the principle of explicability points out the important of information, communication, traceability and transparency of the information of AI systems and their capabilities. Its statement is as follows:

"Explicability is crucial for building and maintaining users' trust in AI systems. This means that processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected. Without such information, a decision cannot be duly contested. An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible. These cases are referred to as 'black box' algorithms and require special attention. In those circumstances, other explicability measures (e.g. traceability, auditability and transparent communication on system capabilities) may be required, provided that the system as a whole respects fundamental rights. The degree to which explicability is needed is highly dependent on the context and the severity of the consequences if that output is erroneous or otherwise inaccurate." (AI-HLG, 2019b, 15)

As observed, these principles provide important guidance on the development of new regulation, particularly in anticipating possible harms associated with the development, deployment and use of AI in EU. At the same time, they provide ethical guidelines to take into account during these three phases of AI implementation systems. This second aspect is development in the next layer of the proposal. Potential conflicts could arise in situations in which you have to decide whose principles or principle you have to prioritize. The heuristic here is that "(c)ertain fundamental rights and correlated principles are absolute and cannot be subject to a balancing exercise (e.g. human dignity)."(AI-HLG, 2019b, 15).

2) **Realising Trustworthy AI**. This is the most detailed part of the proposal. It is divided into two sections. The first section introduces in detail the seven requirements for implementing AI. The second, discuss technical and not technical methods for achieving these requirements.

Requirements

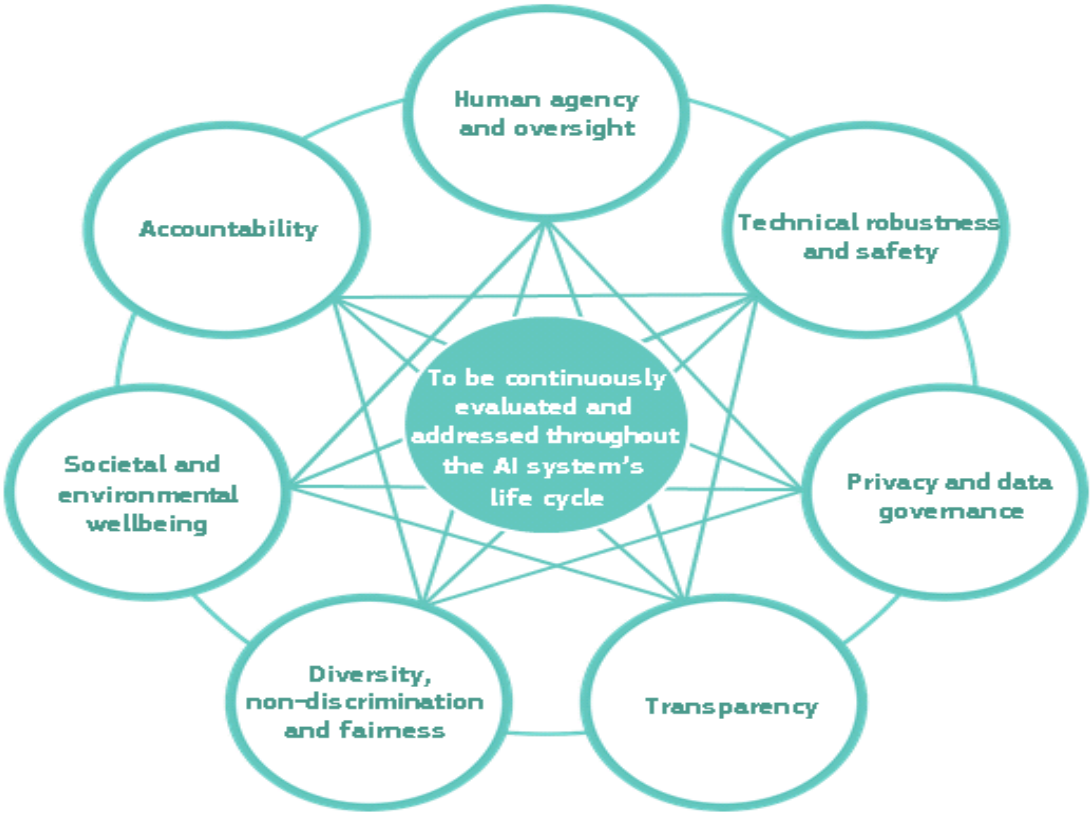
As recalled, seven requirements informed this proposal for realizing trustworthy AI. These are:

- "Human agency and oversight,
- Technical robustness and safety,

- Privacy and data governance,
- Transparency,
- Diversity, non-discrimination and fairness,
- Societal and environmental wellbeing, and
- Accountability"

The relationships between these requirements are represented in figure 1. According to this high level group, these requirements have to be complied in all the stages and during all the entire life-cycle of AI systems. And they should be adapted to the specific area of application of AI systems, determining the level of strictness required. As the AI-HLG mention, AI systems for manufacturing process could be less strict that those systems that process personal data.

Figure 1
The system of requirements



Source: AI-HLG, 2019b, 17

On human agency and oversight three main issues are considered. A) The way in which AI can use to "enable and hamper fundamental rights" making easier for people to track his personal data, to facilitate the access to education or health services. Particular attention should be payed to the protection of the individual, the strengthen of democracy, the rights and freedom of people, preventing, at the same time, those situations that can threaten them. B) The human agency should be preserved. This means that people should have the tools and knowledge needed to understand the benefits, but also the risks, associated with AI applications and the preventions implemented to reduce or eliminate them. "Key to this is the right not to be subject to a decision based solely on automated processing when this produces legal effects on users or similarly significantly affects them." (AI-HLG, 2019b, 18). C) The Human oversight is very important to verify that the principles that inform the development, deployment and use of AI systems are satisfied. Among the mechanisms envisaged, three design mechanisms are important to mention:

"Oversight may be achieved through governance mechanisms such as a human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approach. HITL refers to the capability for human intervention in every decision cycle of the system, which in many cases is neither possible nor desirable. HOTL refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation. HIC refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation" (AI-HLG, 2019b, 18).

The requirement "Technical robustness and safety" is connected to trustworthy principle of preventing from harm and deals with the following four aspects: A) *Resilience to attack and security* in which the protection against vulnerabilities under the form of data poisoning and model leakage and "underlying infrastructure, both software and hardware" should be assured. Tracking changes system, safety redundancy, procedures for potential malicious actors' identification and verification, and strategies for limiting and mitigating these impacts are included among the measures needed. B) *Fallback plan and general safety*includes safeguards such as statistical to rule-based procedures or asking-human operator before continuing acting, processes for clarification potential risks in specific applications, among others. As mentioned before, the strictness of these procedures depend on the application area of these systems. C) Accuracy. "Accuracy pertains to an AI system's ability to make correct judgements, for example to correctly classify information into the proper categories, or its ability to make correct predictions, recommendations, or decisions based on data or models". Experience shows that a good design, well developed process and good evaluation contribute importantly to the success of systems and make it easier to identify potential risks. D) Reliability and Reproducibility. These are two consolidated practices in scientific community and it is relevant to apply them strictly to AI systems aimed at assuring the comply with the principles drawn above.

Privacy and data governance is an important requirement to assure the protection of personal data, and tracking in case of intended or unintended harm behavior of AI systems. It is related to the following three issues: A) Privacy and data protection. B) Quality and integrity of data, C) Access to data. These issues have been widely discussed in several forums related to technology, so we will not discuss these in detail here.

Transparency is the four requirement. This requirement relies on two fundamental aspects: Strict standards for documentation and communication; on stakeholders' structure and the clarification of their role in the process of development, deployment and use of AI systems. As introduced in these guidelines, transparency is composed of three processes: Traceability, explainability and communication. In this sense, it is clearly connected with other requirements of the system. Documentation and communication are under the responsibility of developers and those involved on the deployment of AI systems, and verification is responsibility of all the stakeholders involved in all the stages of the life-cycle of the product.

Diversity, non-discrimination and fairness is the fifth requirement. Several professions and perspectives are involved in the achievement of this requirement. According to AI-HLG, diversity should be considered during all the life-cycle of the AI system. And should identify potential risk related to discrimination and unfair; and conveys the following critical aspects: Avoidance of unfair bias, accessibility and universal design, and stakeholder Participation. It is one of the most complex component of this trustworthy system. The development of protocols and guidelines for standardising the information obtained during all the processes of consultation and participation is crucial; but also the way in which this information is transformed into further system´s requirements and system evaluation criteria.

Societal and environmental well-being. This requirement reinforces the need of focus EU efforts to development those AI systems that impact positively on well-being and environment, such as found in coordinated plan discussed above. In addition to these, AI-HLG calls for considering the following three aspects: Sustainable and environmentally friendly AI, social impact (considering both the enhance social skills but also its deterioration due to the use of technology), and society and democracy (as indicated in the principles, to improve society and democracy as well).

Finally, accountability. "It necessitates that mechanisms be put in place to ensure responsibility and accountability for AI systems and their outcomes, both before and after their development, deployment and use"(AI-HLG, 2019b, 21). Accountability conveys three important processes: Auditability, the assessment of algorithms, design processes and deployment; minimisation and reporting of negative impacts aimed at introduce the require corrections (redress) and trade-offs. Trade-offs is a very complex issue that require a

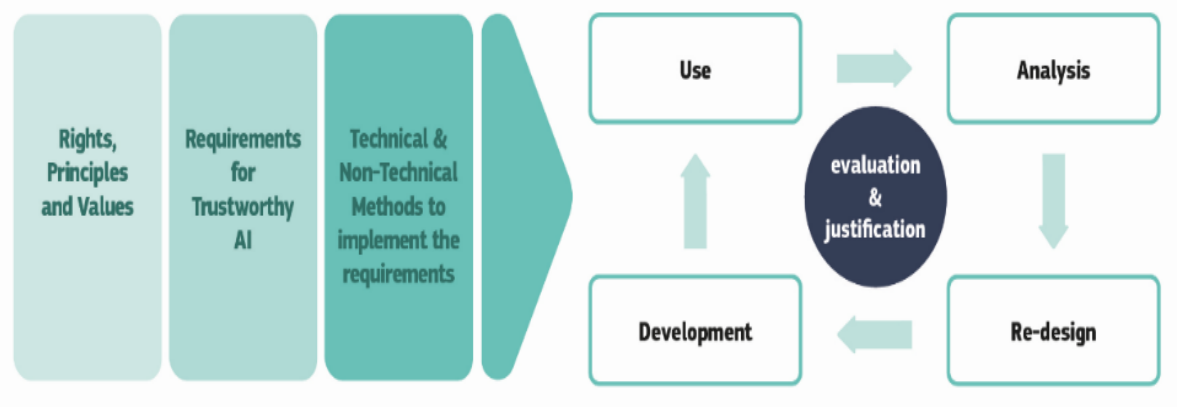
standard and methodic procedure for addressing them, but in all cases, fundamental rights, values and principles should prevail.

Technical and non-technical methods

In order to accomplish these requirements, the use of technical and non-technical methods is needed. Principles and requirements are the more permanent components of the system of trustworthy AI (form the kernel of the system), but some adaptation should be made as progress is reached. The more dynamic part of the system is formed by the technical and non-technical methods. The goal of these is to reach higher levels of compliance with the principles and the requirements. AI-HLG captures these two aspects of the trustworthy AI system in the following flow.

Figure 2

Relationships between principles, requirements and technical and non-technical methods



Source: AI-HLG, 2019b, 22

Several technical methods are proposed. They are arranged according to the level of maturity. What this means is that it is needed more research to improve some of these methods that could play an important role in assuring AI systems and to yield the expected results.

The first technical method is called "Architectures for Trustworthy AI". Requirements should be translated into procedures or constrains anchored into architectures. According to this,

"AI systems with learning capabilities that can dynamically adapt their behaviour can be understood as non-deterministic systems possibly exhibiting unexpected behaviour. These are often considered through the theoretical lens of a "sense-plan-act" cycle. Adapting this architecture to ensure Trustworthy AI requires the requirements' integration at all three steps of the cycle: (i) at the "sense"-step, the system should be developed such that it recognises all environmental elements necessary to ensure adherence to the requirements; (ii) at the "plan"-step, the system should only consider plans that adhere to the requirements; (iii) at the "act"-step, the system's actions should be restricted to behaviours that realise the requirements." (AI-HLG, 2019b, 23)

The second technical method is called "Ethics and rule of law by design (X-by-design)" and consists of identifying from the beginning impacts and also norms that ought to be taking into account in the design of a AI system. Some of these methods are used in the development of computer systems, such as "privacy-by-design and security-by-design". These are proposed to guarantee that the system is secure "in its processes, data and outcomes, and should be designed to be robust to adversarial data and attacks" (AI-HLG, 2019b, 23).

The third proposed technical method is "Explanation methods" (XAI)that consists of the ability of the system to provide explanations of the way in which it behaves and the interpretation it gives to that act. AI-HLG mentions the challenges posed, for example, by neural network in which this explanation and interpretation is hard to obtain. "Methods involving XAI research are vital not only to explain the system'sbehaviour to users, but also to deploy reliable technology".

Testing and validating is the fourth proposed method. This is well consolidated in the practice of computer science, however, AI system works in non-deterministic and context sensitive ambient that makes

necessary to advance in the development of appropriate methods for these systems. These should "verify and validate processing of data" and "the underlying model must be carefully monitored during both training and deployment for its stability, robustness and operation within well-understood and predictable bounds".

Finally, technical methods of Quality of Service Indicators should include "measures to evaluate the testing and training of algorithms as well as traditional software metrics of functionality, performance, usability, reliability, security and maintainability" (AI-HLG, 2019b, 24).

Non-technical methods. AI-HLG considers eight non-technical methods as candidates for achieving trustworthy AI.

The first is regulation. As technology progress new regulations are needed to assure that these technologies be at the service of human being and the environment. Multistakeholder instances can be an important source for proposing regulation´s up-dating.

The second non-technical method is Code of Conduct. These codes are widely used to regulate the practice of different group of professionals. And this could play an important role in reinforcing conducts and practices to achieve trustworthy AI.

The third method standardisation. It is widely acknowledged the relevance of systems of norms such as ISO, IEEE or ACM, to mention only three, in the promotion of good and standardised professional practices. In the same vein, this kind of standardisation could be relevant for implementing this trustworthy AI strategy.

The fourth proposed method is certification. Certifications complements standardisation, in the sense that the last also can be the object of certification. But in general,

"(t)hese certifications would apply standards developed for different application domains and AI techniques, appropriately aligned with the industrial and societal standards of different contexts. Certification can however never replace responsibility. It should hence be complemented by accountability frameworks, including disclaimers as well as review and redress mechanisms" (AI-HLG, 2019b, 25).

Four more methods are mentioned in these Guidelines: Accountability via governance frameworks that includes, for example, boards of ethical experts for evaluation and justification. Education and awareness to foster an ethical mind-set, that could include diverse activities in which stakeholders and other professionals participate in discussions on the issue and in proposing ideas to improve the implementation of the system. Stakeholder participation and social dialogue could contribute to discuss the issue in different social stratus and to channel suggestions to improve the system. Finally, diversity and inclusive design teams could also play an important role in this communitarian enterprise.

IV. Assessing Trustworthy AI

This section includes a very interesting number of suggestions on how to operationalise trustworthy AI proposal. It explicitly excluded two issues. The first one is how AI systems have to comply with law and regulations. To do this it is needed an additional work that includes an analysis of law, the foundational charter of EU and the international treaties on human rights. The second issue concerns the operationalisation of the requirements from the end-user perspective. As we have seen, two aspects should be mentioned on it: a) the issue hides enormous complexity due to diversity of backgrounds and perspectives of end-users and end-user stakeholders; b) it remains as future work to develop in details non-technical methods to use in the evaluation and verification of AI systems.

So the operationalization of EU AI proposal deals with two main components of the proposal: Development of and deploymentAI systems. The section is organized into two parts: one short part on the multistakeholder structure proposed for the development of and deploymentAI systems, and the second part on checking list suggestions to comply with principles and requirement.

We will do a very brief reference to each one. On the multistakeholder structure, it is composed by the following relevant levels:

- Management and Board,
- Compliance/Legal department/Corporate responsibility department
- Product and Service Development or equivalent
- Quality Assurance
- Human Resources

- Procurement
- Day-to-day Operations

It is briefly described the role of each of stakeholder level in development and deployment of AI systems.

On the checking list part, is a requirement approach in which several questions need to be answered during the development and deployment of these systems. The reader will find them interesting and useful to understand from a practical perspective the implementation of these guidelines.

Finally, the Guidelines concludes with considering "examples of opportunities and critical concerns raised by AI". Several opportunities for the development of these systems are presented, but also some critical concerns raised by these systems. Here also, the reader will find interesting cases for discussion.

V. Concluding remarks

We have presented in detail EU proposal for a trustworthy AI because it could be an important starting point for regulating AI development and deployment in other countries, like ours. As shown, potentialities and also concerns arise from AI. The benefits of this technology surpasses concerns when appropriately approached. We are entering into a new massive data era and it is proved that AI algorithms have better performance working on big data. These systems can be used to infer new information from data that can help in the process of decision making, but also can be used to invade privacy and to commit cybernetic crimes, among others. The best way to optimize benefits and reduce risks is constructing an appropriate framework that allows us to understand, evaluate and verify that these systems will contribute to enhance human capabilities and to improve our environment.

References

[1]. Acemoglu and Restrepo, 2019 The Wrong Kind of AI? Artificial Intelligence and the Future of Labor Demand. URL =<https://www.nber.org/papers/w25682>

[2]. AI-HLG (2019b) Ethics Guidelines for Trustworthy AI. URL = <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>

[3]. AI-HLG (2019a) A definition of Artificial Intelligence: main capabilities and scientific disciplines. URL= <https://ec.europa.eu/knowledge4policy/online-resource/definition-artificial-intelligence-main-capabilities-scientific-disciplines_en>

[4]. Bringsjord, Selmer and Govindarajulu, Naveen Sundar, "Artificial Intelligence", The Stanford Encyclopedia of Philosophy (Summer 2020 Edition), Edward N. Zalta (ed.), forthcoming URL = <https://plato.stanford.edu/archives/sum2020/entries/artificial-intelligence/>.

[5]. Buckner, Cameron and Garson, James, "Connectionism", The Stanford Encyclopedia of Philosophy (Fall 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2019/entries/connectionism/>.

[6]. National Science & Technology Council, 2019, The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update. URL =<https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf>

[7]. Russell, S. &Norvig, P., 2009, Artificial Intelligence: A Modern Approach 3rd edition, Saddle River, NJ: Prentice Hall.

[8]. Thomason, Richmond, "Logic and Artificial Intelligence", The Stanford Encyclopedia of Philosophy (Summer 2020 Edition), Edward N. Zalta (ed.), forthcoming URL = <https://plato.stanford.edu/archives/sum2020/entries/logic-ai/>.

[9]. WEF, 2019, A Framework for Developing a National Artificial Intelligence Strategy. URL= <http://www3.weforum.org/docs/WEF_National_AI_Strategy.pdf>

[10]. The European Union, 2020b, WHITE PAPER On Artificial Intelligence - A European approach to excellence and trust. URL = <https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en>

[11]. The European Union, 2020a Communication Coordinated Plan on Artificial Intelligence (COM(2018) 795 final). URL = <https://ec.europa.eu/knowledge4policy/publication/coordinated-plan-artificial-intelligence-com2018-795-final_en>