The protection of Human Rights using Artificial Intelligence to fight against inequalities

EXECUTIVE SUMMARY

Someday machines will be able to solve all problems, but never will any of them be able to pose one (*Albert Einstein*).

If progress cannot be stopped, it is fundamental to understand the **mechanisms of technologies** to be able to **intervene, prevent** and secure the delicate balance between artificial intelligence and the protection of human rights.

Today we speak of pervasive computing, or the incorporation of microprocessors into everyday objects, enabling them to communicate information, so that virtually any device, from clothing to kitchen appliances, can be connected thank to microchips to unlimited network of other devices.

Well, technology already touches the everyday life of the contemporary world, highlighting the limits and dangers of being perpetually surrounded by technology and raising awareness of the moral issue related to it.

In this paper we are going to analyze the relevant **relationship between artificial intelligence and the right to life**, and particularly the **right to health** protection and **how artificial intelligence can help to improve psychophysical well-being**.

Nowadays and especially after the COVID-19 outbreak, the psychophysical well-being must be protected as it is the biggest workforce problem of our time and will be for the next decade.

Time has come for organizations to explore solutions to address it.

The starting point is elaborating an adequate legal framework that will ensure the proper attention to the new ethical and legal questions that Artificial Intelligence – as any other transformative technology – raises, while ensuring the promotion of innovation.

What we would like to show is that the **application of technologies can be used as a booster to improve the quality of** life of people focusing on **employees**, while overcoming diversities and helping disadvantaged individuals, such as the one with disabilities, to find a better work and life condition and balance.

There are practical tools that may allow companies to address the challenges of **filling the gaps of the inequalities** in the human rights protection. This paper proposes solutions that can be implemented by multinational companies in order to be the first mover and create a positive trend and being the *influencer* of other corporation.

Human capital must be enhanced and protected also with Artificial Intelligence.

Technology is not the end but the means.

The protection of Human Rights using Artificial Intelligence to fight against inequalities

Summary

Authors

Valentina Rispoli

Andrea Brun

Serena Corbetta

Adriano Conte

Marianna Dolcetti

Tutor

Ludovica Maria Vittoria Parodi Borgia

1. Introduction

The right to health, the right to privacy, the right to a fair trial, the freedom of expression, the right of access to public services and health care, the right to work, are all fundamental rights that are to be protected and discrimination is to be avoided.

Recent advances in robotics and Machine Learning are enabling the development of systems that can compete with human capabilities in specific areas or tasks while also allowing, in some cases, to surpass them.

Thanks to Machine Learning, or automated learning, these systems are able to learn from experience and mistakes made, improving exponentially in terms of efficiency and independence.

These advances are enabling the increasing use of such systems in various fields such as, inter alia, industrial production, healthcare, public assistance or the legal sector.

The progressive and extensive use of these systems, however, if on one hand stands as a new frontier of innovation, on the other hand could raise problems where there is not, for example, a proper process of adaptation and implementation in the current regulatory and organizational system.

The growing and deepening relationship between people and technology contributed to questioning about issues correlated to the design methods of technological innovations. This implicitly confers an important social and ethical role to the software houses and companies as Enel X, which contribute to the creation of products that can influence the behavior of the users to the point of manipulating it.

However, this same important role also represents a powerful means to promote positive behaviors.

The designers of Artificial Intelligence should consider the impact of their work on people life. With this in mind, an "*ethical design*" movement was born, aimed at promoting digital well-being and revising the methods of technology implementation in a human-centric fashion.

According to a new study by Oracle[1] the global COVID-19 pandemic has exacerbated psychological problems in the workplace. The impact of this pandemic has affected negatively on people's life for various aspects, included professional life. People are feeling the effects in their private lives as well. 85% of people worldwide say that work-related problems with mental health and well-being affect their private lives. Despite some perceived disadvantages to working remotely, 62% of people find remote work more appealing now than before the pandemic due to the possibility to spend more time with family (51%), to rest (31%), and to complete family tasks (30%).

[1] Studio AI @ Work 2020 - Permanere dell'incertezza che porta a un punto di svolta dei livelli di ansia e stress sul lavoro: https://www.oracle.com/a/ocom/docs/oracle-hcm-ai-at-work-italy.pdf

This implies that an increasing number of people spend a considerable part of their time interacting with digital devices, using many services whose use is more or less consciously included in their daily habits.

Social networks, applications for productivity and leisure, fitness and e-commerce, are software commonly used by millions of individuals and whose success often coincides with being carefully designed to make their use a habit. In fact, countless digital products are constantly competing for the user's attention to become part of their routine and to be used as frequently as possible.

Technology, if properly conceived, is in fact a means capable of altering user behavior in a silent but at the same time profound way, to the point of inducing real habits.

The resulting social power can be used to stimulate users to both positive and negative behaviors.

Artificial Intelligence could help humans to stop smoking or lose weight, as well as being used for the sole purpose of monetizing their attention, with the risk of leading them to compulsive behavior or even pathological addiction.

The outcome depends largely on the intentions of the designers of algorithms and the awareness of the users.

Technology persistently continues its rapid democratization process, becoming accessible to more and more people. Reduced costs and a greater focus on user experience have, in fact, encouraged access to technologically sophisticated products and services. However, although on one hand this phenomenon has brought countless economic and social benefits in recent decades, on the other the growth in the general understanding of how it works and its potential risks is not keeping pace with the speed at which the relationship between the individual and the digital device is intensifying.

Before analyzing the application of the artificial intelligence to solve or mitigate the inequalities we should briefly illustrate what artificial intelligence is and, above all, its best-known applications Machine Learning and Deep Learning.

The dynamics within society, at all levels, are profoundly affected by digital transformation. We are living in a new era, the digital age, and every expression or relationship must be completely revised considering the new model.

At the heart of this paradigm, however, the most important factor remains the human one.

2. Artificial Intelligence

2.1 Definition Machine Learning and Deep Learning

Artificial Intelligence represents one of the topics of greatest interest in the current historical period and its best-known applications are Machine Learning and Deep Learning.

There are numerous definitions of Artificial Intelligence. However, we will rely on the following one which defines Artificial Intelligence as "*systems that exhibit intelligent behavior, analyzing the surrounding environment and performing specific actions - with a certain degree of autonomy - to achieve specific results*"[2].

Machine Learning algorithms have been developed since the 1960s, but, due to the limited computing power of the computers of that time, they were mainly applied in the academic field.

Nowadays, however, thanks to the accessibility of machines with high computing power, Machine Learning has aroused renewed interest and has been applied with success in numerous widely used applications.

In addition to that, the diffusion of the so-called big data has done nothing but give further impetus to applied research in the field of Machine Learning, given that these algorithms require significant amounts of data to be analyzed. This is because the machine must be able to identify rules or patterns within the data, in order to be able to take the best decision in future cases.

There are numerous Machine Learning algorithms which, usually, are reduced to three macro categories.

A. **Supervised Learning algorithms**. They are characterized by the fact that a fully annotated dataset is supplied to the machine, both with reference to the input data and with reference to the output data. A simple example could be a system for classifying emails as “spam”. The dataset will, in fact, consist of a series of examples of emails, some of which are classified as "spam" and others classified as "not spam". In the training phase, the algorithm will try to "learn" a function that can identify the expected output for a given input. For example, the inbox of e-mails is automatically divided into two macro categories: spam and not spam and is trained to identify and divide the incoming email.

B. **Unsupervised Learning algorithms**. They are characterized by the fact that the dataset is not annotated. The algorithm will identify patterns, similarities or differences in the data provided for training. A typical example of unsupervised learning is Latent Dirichlet Allocation (LDA), a Machine Learning algorithm widely used in the Natural Language Processing sector that can be used to identify the subject of a portion of text. In this case, the dataset could consist of a set of newspaper articles, not previously classified, and the LDA algorithm could be used to divide the articles into categories, based on the subject matter. For example, photos take with mobile phone are more and more stored by categories and/or subject such as sports, mountains, cooking etc.

[2] European Commission, Communication on Artificial Intelligence for Europe, 2008.

C. **Reinforcement Learning Algorithms**. The machine is basically called upon to make decisions by interacting with the surrounding environment, learning from its mistakes and past experiences. The core of these algorithms is the reward awarded in the event that the action is deemed correct. A reinforcement learning application can be autonomous driving due to strong interactions with other vehicles, pedestrians and roadworks or traffic lights.

Having said that on the categories of Machine Learning algorithms, it is legitimate to ask what Deep Learning is.

The Deep Learning, a sub-category of Machine Learning, it creates learning models on multiple levels. Scientifically, it is correct to define the action of Deep Learning as the learning of data that are not provided by humans but are learned thanks to the use of statistical calculation algorithms. These algorithms aims to act as the human brain interpreting images and language. The learning thus achieved has the shape of a pyramid: the highest concepts are learned starting from the lowest levels.

Deep Learning has achieved remarkable results that once seemed unattainable. For the continuous evolution of artificial intelligence, an incessant amount of data is required to allow the computer to experience and learn. Deep learning plays a fundamental role as it gives the representation of the data, but at a hierarchical level by processing the different levels. This transformation is amazing because it allows us to witness a machine that is able to classify incoming (input) and outgoing (output) data, highlighting the important ones. The revolution brought about by Deep earning is all in the human-like ability to process data, one's knowledge at levels that are not linear at all. Thanks to this faculty, the machine learns and improves ever more complex functions.

Development of AI system are detailed under Appendix 1.

2.2 AI for the public interest

2.2.1 AI minimum requirements

The literature developed on the point has found three main characteristics for an AI to be positively evaluated:

- ***No-discrimination or Impartiality***
  In order to ensure that the single AI tool will work properly, a preliminary check needs to be done on the data that such tool will be going to manipulate: the sets of data to be used need to be wide enough and complete (as much as possible) to be considered meaningful, they should be disaggregated when needed (taking into consideration different sub categories, if any), they should be selected, pre-screened, "prepared" (according to objective criteria established upfront[3]).

[3] *The Achilles' Heel Of AI*, Ron Schmelzer, 2019, Forbes, available at: https://www.forbes.com/sites/cognitiveworld/2019/03/07/the-achilles-heel-of-ai/?sh=21e4dc9a7be7. According to this article, the minimum activities to be conducted while preparing data sets are the following: 1. Removing or correcting bad data and duplicates. 2. Standardizing and formatting data. 3. Updating out of date information. 4. Enhancing and augmenting data. 5. Reducing noise. 6. Anonymizing and de-biasing data. 7. Normalizing data. 8. Preparing data samples. 9. Enhancing features.

AI development shall be enlightened by the basic principle of computer science: avoiding the "GIGO effect" (garbage in – garbage out)[4] [5]. AI developers and users need to be well aware that the quality of the output produced by their futuristic tools is directly proportional to the quality of the input given to the machine: in fact, in every processing system, the data's quality coming out cannot be better than what went in[6]. Or, in other words, a program will only yield misleading results if it is working on faulty data[7].

But the quality of the output also depends on the very way of functioning of the algorithms themselves: they shall be developed avoiding the well-known "**bias of the programmer**"[8] which inevitably leads to the "bias of the algorithm"[9]. Eliminating such kind of bias is not an easy task: even if there are no quick fixes, it is nonetheless possible to improve the quality of algorithms understanding and measuring their "fairness"[10] by "weighting" the output on the basis of indexes such as algorithms' "accuracy", "sensitivity", "performance degradation", "fallacy"[11].

After the interpolation has been done (possibly as described above), a similar process of "attributing significance" to the data output should be conducted by someone having a solid knowledge of the specific matter with the scope of "reading" the data; and, again, outputs should be disaggregated (if possible, if necessary, if helpful) and considered in the wider context that they represent and where they will be used[12].

This is called "**independent validation**" of the data and it is absolutely essential before using new and high-impact AI systems[13].

- ***Transparency or Explainability***

  As already mentioned in the opening paragraph of this paper, "transparency" is a prerequisite for an algorithm to be considered "trustworthy": after any interpolation of data, results should be explicable and this could be difficult, if not impossible, for algorithm using multiple layers. Despite the inevitable trade-off between model complexity and

[4] *Avoiding Garbage in Machine Learning*, S&P Global, 2019, available at: https://www.spglobal.com/en/research-insights/articles/avoiding-garbage-in-machine-learning-shell.

[5] *AI And Machine Learning In Healthcare: Garbage In, Garbage Out,* Jeff Gorke, Forbes, 2019, available at: https://www.forbes.com/sites/jeffgorke/2020/06/18/ai-and-machine-learning-in-healthcare-garbage-in-garbage-out/?sh=4d252ff750a7.

[6] *What is GIGO (garbage in, garbage out)?*, MBN – Market Business News, available at: https://marketbusinessnews.com/financial-glossary/gigo-garbage-in-garbage-out/

[7] *A guide to healthy skepticism of artificial intelligence and coronavirus*, Alex Engler, Brookings, 2020, available at: https://www.brookings.edu/research/a-guide-to-healthy-skepticism-of-artificial-intelligence-and-coronavirus/

[8] *Don't blame the AI, it's the humans who are biased*, Samara J Donald, 2019, available at: https://towardsdatascience.com/dont-blame-the-ai-it-s-the-humans-who-are-biased-d01a3b876d58

[9] *Id*.

[10] *What Do We Do About the Biases in AI?*, James Manyika, Jake Silberg, and Brittany Presten, Harvard Business Review, 2019, available at: https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai

[11] *A guide to healthy skepticism of artificial intelligence and coronavirus*, cit.

[12] On the concept of "meaningful information", see *#BigData: discrimination in data-supported decision making*, FRA – EUROPEAN UNION AGENCY FOR FUNDAMENTAL RIGHTS, 2018, available at https://fra.europa.eu/en/publication/2018/bigdata-discrimination-data-supported-decision-making.

[13] *A guide to healthy skepticism of artificial intelligence and coronavirus*, cit.: "*No matter what the topic, AI is only helpful when applied judiciously by subject-matter experts—people with long-standing experience with the problem that they are trying to solve. Despite all the talk of algorithms and big data, deciding what to predict and how to frame those predictions is frequently the most challenging aspect of applying AI. Effectively predicting a badly defined problem is worse than doing nothing at all. Likewise, it always requires subject matter expertise to know if models will continue to work in the future, be accurate on different populations, and enable meaningful interventions*."

interpretability, the functioning of AI tools cannot be totally obscure: it shall be possible, albeit complex, to reconstruct the reasoning followed by the machine.

In this context, transparency, explainability or explicability are all synonyms of the general concept of "interpretability of results": where "interpretable" stands for possibility to comprehend the influencing factors of the decisions taken by specific AI tools applied, with the scope of verifying their lawfulness, and, in case, of intervening with corrective actions.

- ***Accountability***

  Third character of a "safe" AI is accountability: being transparent, in fact, does not necessarily help the proper use of this powerful technology, at least, if not accompanied by a chain of accountability that holds the systems human operator responsible for the decisions of the algorithm.

  In terms of regulation, accountability means also the ability to determine whether a decision was made in accordance with procedural and substantive standards and to hold someone responsible if those standards are not met[14].

  For an AI to be acceptable, then, there must be the possibility to explain and justify one's decisions and actions. To ensure accountability, decisions must be derivable from, and explained by, the decision-making algorithms used. This includes the need for representation of the moral values and societal norms holding in the context of operation, which the agent uses for deliberation. Accountability in AI requires both the function of guiding action (by forming beliefs and making decisions), and the function of explanation (by placing decisions in a broader context and by classifying them along moral values)[15].

2.2.2 AI and human rights

The evaluation of the AI in relation with human rights is an activity that has caught the general public completely unprepared.

In a number of interviews conducted by the European Union Agency for Fundamental Rights ("FRA") among more than a hundred stakeholders (private companies, but also public offices), a dramatically high percentage of the interviewees declared that the specific tool of AI that they were commenting had none or limited potentially negative impacts on any fundamental right[16].

Such answers, although given in good faith, were all proven wrong once giving a broader look to the contexts in which each AI tool was applied. In fact, following a more accurate exam of each case, negative outcomes were found with reference to a number of fundamental rights, particularly to the right of privacy, freedom of speech, freedom of association, just to name a few.

The reasons of such misalignment between the "*real impacts*" and the "*perceived impacts*" of the AI use are to be found in the common tendency of considering only or preferring the positive outcomes – which are immediately "*perceived*" – of each concrete application (such as, for

[14] *Accountability of AI Under the Law: The Role of Explanation,* VV.AA., Berkman Klein Center Working Group on AI Interpretability, Working Draft, available at: https://arxiv.org/ftp/arxiv/papers/1711/1711.01134.pdf

[15] The ART of AI — Accountability, Responsibility, Transparency, Virginia Dignum, Medium.com, 2018, available at: https://medium.com/@virginiadignum/the-art-of-ai-accountability-responsibility-transparency-48666ec92ea5

[16] Research methodology available here: https://fra.europa.eu/sites/default/files/fra_uploads/fra-2020-artificial-intelligence-annex-1_en.pdf

example, the very automation of certain functions, the reduction of time to complete a task, the increased level of precision of certain outcomes, the possibility to process a greater number of data than could be done manually, just to name a few), neglecting the negative outcomes – which, instead, are more difficult to spot, as less eye catching, and require a more sophisticated exam to be caught. In addition, it is clear that the lack of knowledge and awareness of the potential negative implications of the use of AI in the contiguous contexts also derives from the novelty of the technology itself as well as from the consequent reduced number of case studies available.

Having focused this short-circuit, the FRA[17], with the aim of correcting such misperception and preventing the misapplication of the AI, recommended – for each new AI tool to be proposed – the preventive performance of a "Fundamental Rights Impact assessment" ("FRIA"), having the specific goal of discovering all the possible encroaches between the AI tool use and all the possible fundamental rights[18].

According to the suggested approach, the study of the potential interferences of AI with other "protected areas" is of pivotal importance to ensure that the application of AI will benefit the common good, ensuring results that be exponentially positive, with so avoiding to become an evil booster exacerbating existing inequalities, or the root of new inequalities or, again, the origin of

[17] G*etting The Future Right Artificial Intelligence And Fundamental Rights*, European Union Agency for the Fundamental Rights, 2020, available at: https://fra.europa.eu/sites/default/files/fra_uploads/fra-2020-artificial-intelligence_en.pdf

[18] To do so, a list of possible questions is given (drawing on specific articles in the Charter and the European Convention on Human Rights ("ECHR") its protocols and the European Social Charter), but of course, users should develop their own list, that should demonstrate that a sufficient investigation has been conducted on the point prior of developing the specific tool. Examples of questions – available at file:///C:/Users/E351252/Downloads/altai_final_14072020_cs_accessible2_jsd5pdf_correct-title_3AC24743-DE11-0B7C-7C891D1484944E0A_68342%20(1).pdf are: "*1. Does the AI system potentially negatively discriminate against people on the basis of any of the following grounds (non-exhaustively): sex, race, color, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation? Have you put in place processes to test and monitor for potential negative discrimination (bias) during the development, deployment and use phases of the AI system? Have you put in place processes to address and rectify for potential negative discrimination (bias) in the AI system? 2. Does the AI system respect the rights of the child, for example with respect to child protection and taking the child's best interests into account? Have you put in place processes to address and rectify for potential harm to children by the AI system? Have you put in place processes to test and monitor for potential harm to children during the development, deployment and use phases of the AI system? 3. Does the AI system protect personal data relating to individuals in line with GDPR?16 Have you put in place processes to assess in detail the need for a data protection impact assessment, including an assessment of the necessity and proportionality of the processing operations in relation to their purpose, with respect to the development, deployment and use phases of the AI system? Have you put in place measures envisaged to address the risks, including safeguards, security measures and mechanisms to ensure the protection of personal data with respect to the development, deployment and use phases of the AI system? See the section on Privacy and Data Governance in this Assessment List, and available guidance from the European Data Protection Supervisor.17 4. Does the AI system respect the freedom of expression and information and/or freedom of assembly and association? Have you put in place processes to test and monitor for potential infringement on freedom of expression and information, and/or freedom of assembly and association, during the development, deployment and use phases of the AI system? Have you put in place processes to address and rectify for potential infringement on freedom of expression and information, and/or freedom of assembly and association, in the AI system?*".

other worrisome issues[19][20]. The tools to conduct this analysis, can and should be found in the international human rights arena[21], as long as the analytical techniques developed in such field can help the identification and anticipation of possible social harms deriving from the indiscriminate use of AI and guide the development of technical and policy safeguards to promote instead its positive uses[22].

The starting point is elaborating an adequate legal framework that will ensure the proper attention to the new ethical and legal questions that AI – as any other transformative technology – raises, while ensuring the promotion of innovation. The debate, therefore, should be led by the EU, followed by the member States, called to implement at national level the general principles adopted at the European level,[23] but it would also involve other important stakeholders such as the global human rights community, and the UN bodies, the domestic human rights institutions

[19] *Human rights in the age of artificial intelligence*¸ Study, AccessNow.org, available at: https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf: "*Many of the issues that arise in examinations of this area are not new, but they are greatly exacerbated by the scale, proliferation, and real-life impact that artificial intelligence facilitates. Because of this, the potential of artificial intelligence to both help and harm people is much greater than from technologies that came before*".

[20] *Does AI stands for augmented inequality in the era of COVID-19 healthcare?*, BMJ 2021;372:n304, http://dx.doi.org/10.1136/bmj.n304, visually representing the Cascade effect of health inequality and discrimination manifest in the design and use of artificial intelligence (AI) systems:

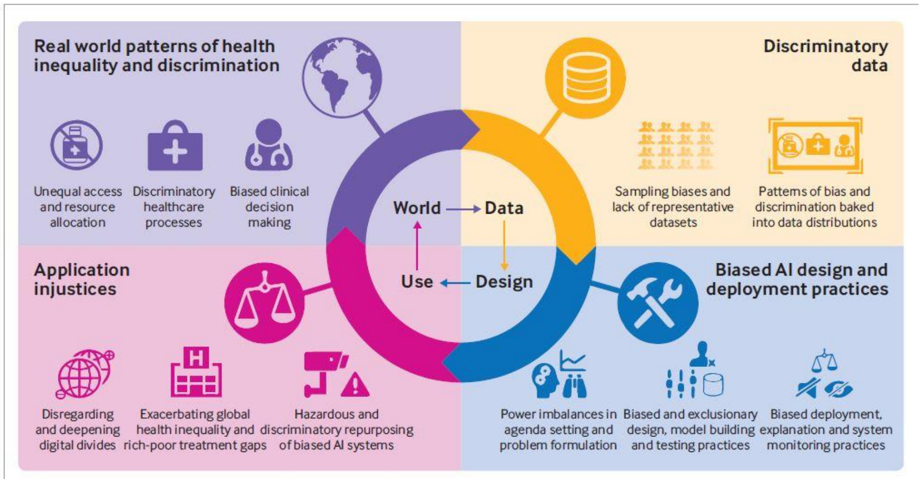

Real world patterns of health inequality and discrimination
Unequal access and resource allocation
Discriminatory healthcare processes
Biased clinical decision making
Discriminatory data
Sampling biases and lack of representative datasets
Patterns of bias and discrimination baked into data distributions
World
Data
Use
Design
Application injustices
Disregarding and deepening digital divides
Exacerbating global health inequality and rich-poor treatment gaps
Hazardous and discriminatory repurposing of biased AI systems
Biased AI design and deployment practices
Power imbalances in agenda setting and problem formulation
Biased and exclusionary design, model building and testing practices
Biased deployment, explanation and system monitoring practices

Fig 1 | Cascading effects of health inequality and discrimination manifest in the design and use of artificial intelligence (AI) systems

[21] *Governing Artificial Intelligence: Upholding Human Rights & Dignity*, Mark Latonero, Data & Society, available at: DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights: "*In order for AI to benefit the common good, at the very least its design and deployment should avoid harms to fundamental human values. International human rights provide a robust and global formulation of those values*".

[22] Id.: "*Implementing human rights can help identify and anticipate some of AI's worst social harms and guide those developing technical and policy safeguards to promote positive uses.*"

[23] *Artificial Intelligence for Europe*, Communication From The Commission To The European Parliament, The European Council, The Council, The European Economic And Social Committee And The Committee Of The Regions, 2018, available at file:///C:/Users/E351252/Downloads/com2018237_DA6A9B35-FB19-8BFE-A6F51E2EA6BA900B_51625%20(1).pdf: "*As with any transformative technology, some AI applications may raise new ethical and legal questions, for example related to liability or potentially biased decision-making. The EU must therefore ensure that AI is developed and applied in an appropriate framework which promotes innovation and respects the Union's values and fundamental rights as well as ethical principles such as accountability and transparency. The EU is also well placed to lead this debate on the global stage*".

and the human rights NGOs. The overall idea should be ensuring that human rights become part and parcel of discussions on the future of AI[24].

The FRIA will then be followed by the assessment of the specific AI characteristics according to the Assessment List for Trustworthy AI ("ALTAI")[25] which is a list of requirements to be read in combination with the general requirements recall above, that has been introduced to help assess whether the specific AI system that is being developed, deployed, procured or used is "trustworthy".

The seven key requirements that, according to such Guidelines, characterize what is called "Trustworthy Artificial Intelligence" are:

1. **Human agency and oversight**: AI systems should empower human beings, allowing them to make informed decisions and fostering their fundamental rights. At the same time, proper oversight mechanisms need to be ensured, which can be achieved through human-in-the-loop, human-on-the-loop, and human-in-command approaches
2. **Technical Robustness and safety**: AI systems need to be resilient and secure. They need to be safe, ensuring a fall back plan in case something goes wrong, as well as being accurate, reliable and reproducible. That is the only way to ensure that also unintentional harm can be minimized and prevented.
3. **Privacy and data governance**: besides ensuring full respect for privacy and data protection, adequate data governance mechanisms must also be ensured, taking into account the quality and integrity of the data, and ensuring legitimized access to data.
4. **Transparency**: the data, system and AI business models should be transparent. Traceability mechanisms can help achieving this. Moreover, AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned. Humans need to be aware that they are interacting with an AI system, and must be informed of the system's capabilities and limitations.
5. **Diversity, no-discrimination and fairness**: Unfair bias must be avoided, as it could have multiple negative implications, from the marginalization of vulnerable groups, to the exacerbation of prejudice and discrimination. Fostering diversity, AI systems should be accessible to all, regardless of any disability, and involve relevant stakeholders throughout their entire life circle.
6. **Societal and environmental well-being**: AI systems should benefit all human beings, including future generations. It must hence be ensured that they are sustainable and environmentally friendly. Moreover, they should take into account the environment, including other living beings, and their social and societal impact should be carefully considered.
7. **Accountability**: Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes. Auditability, which enables the assessment of algorithms, data and design processes plays a key role therein,

[24] *Artificial Intelligence: What's Human Rights Got To Do With It?*, Christiaan van Veen, Data & Society, 2018, available at Artificial Intelligence_ What's Human Rights Got To Do With It_ _ by Christiaan van Veen _ Data & Society_ Points.
[25] *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment*, Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission available at: https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence.

> especially in critical applications. Moreover, adequate an accessible redress should be ensured. [26]

Far from being an obstacle to the AI further development and implementation, such systematic approach would foster the use of AI by contributing to overcome the extremely high reputational costs that such new technology is currently paying for being perceived as a "human rights violator"[27.]

In chapter 4 that follows, we will apply the described methodology to the AI tool that we are suggesting overcoming inequalities in the field of psychological well-being.

[26] At the following link: https://altai.insight-centre.org/ it is also possible to use the practical tool made available for users.

[27] *Artificial Intelligence: What's Human Rights Got To Do With It?*, cit.

3. SDG health

3.1 Psychophysical well-being and inequalities

Now that we have a clearer idea of what AI means and of how it shall be designed to be considered as “AI for the public good” and “trustworthy”, being also finally aware of its many possible interferences with fundamental human rights, is time to draw the bridge between AI and health.

Having decided to focus this paper on the “business context”, the concept of “health” will be analyzed keeping in mind our target, that is – at this stage- the “employees”, and discussing the specific component of health which is the psychological (or mental) one.

The described approach wants to offer an innovative reading of the concepts of health and wellbeing, but it finds its roots in the general principles of the EU regulations, as it is shown as follows.

Let’s start from the European Sustainable Development Goals, where SDG 3 is specifically dedicated to Good Health and Well Being, aiming at: “*Ensure healthy lives and promote wellbeing for all at all ages*”.

But what does “health” mean?

Under the Constitution of the World Health Organization, health means “*a state of complete* ***physical, mental and social well-being*** *and not merely the absence of disease or infirmity*”[28].

Similarly, the International Covenant on Economic, Social and Cultural Rights adopts a holistic definition of health when recognizing: “[…] *the right of everyone to the enjoyment of the highest attainable standard* ***of physical and mental health*** […]”[29].

[28] See the premises of the Constitution of the World Health Organization (“WHO”), available at: https://www.who.int/governance/eb/who_constitution_en.pdf

[29] See International Covenant on Economic, Social and Cultural Rights (“ICESCR”), art. 12: “*1. The States Parties to the present Covenant recognize the right of everyone to the enjoyment of the highest attainable standard of physical and mental health. 2. The steps to be taken by the States Parties to the present Covenant to achieve the full realization*

These concepts are strictly linked and interrelated with the one of “organizational wellbeing” which is the “*ability of an organization to promote and maintain the highest levels of* ***physical, psychological, and social wellbeing*** *of its employees, no matter their occupation*” [30].

All this considered, it not only makes sense, but rather, it appears highly advisable that stakeholders, organizations and companies in particular, implement measures and activities fostering the mental health of their employees.

This is true not only in theory, but also in practice, especially considering the recent challenges posed to the mental health by the COVID-19 pandemic.

It is now unanimously agreed that, the **psychophysical well-being must be protected** as it is the biggest workforce problem of our time and will be for the next decade and therefore the time has come for organizations to explore solutions to address it.

This is the outcome of several surveys[31] conducted on employees’ feelings during the pandemic: a worrying percentage of employees, during and after the lockdown, felt weaker and more anxious. These moods produced a negative impact on the psychological well-being of the global workforce, causing in particular more stress (38%), lack of balance between work and private life (35%), loneliness (14%), depression due to lack of social life (25%) and burnout (25%).

Concentrating more on the aspect of the remote working, another set of numbers helps us understanding the magnitude of this issue: 84% of workers worldwide and 76% in Italy said they had faced difficulties in remote work, such as the lack of distinction between personal and work life (41%), mental health problems such as stress and anxiety (33 %) which, for 42% of the sample, precipitated personal productivity and, for 40%, led to less effective and thoughtful decisions. Furthermore, 85% worldwide and 78% of Italians affirm that these problems have also had repercussions on private life with sleep deprivation (40%), poor physical health (35%), reduction of domestic serenity (33%), suffering in family relationships (30%) and isolation from friends (28%)[32].

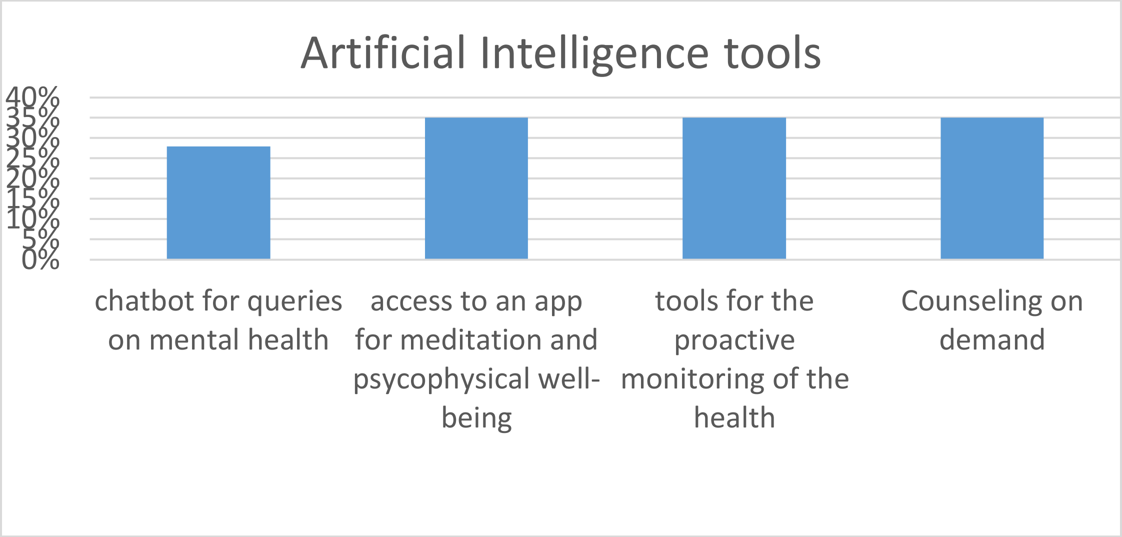
According to the same research, the following graphic identify some **AI tools of which employees would appreciate** the adoption by their company to help them in facing their uncomfortable feelings:

of this right shall include those necessary for: (a) The provision for the reduction of the stillbirth-rate and of infant mortality and for the healthy development of the child; (b) The improvement of all aspects of environmental and industrial hygiene; (c) The prevention, treatment and control of epidemic, endemic, occupational and other diseases; (d) The creation of conditions which would assure to all medical service and medical attention in the event of sickness”.

[30] The concept of “*organizational well-being*” has been firstly developed by prof. Francesco Avallone over the years 2000, and it has been formally introduced in the “*Direttiva della Presidenza del Consiglio 24.3.2004*”, available at https://www.difesa.it/Contatti/URP/Documents/17009_Direttiva_24marzo2004Misure.pdf.

[31] See *AI @work 2020*, Oracle e Workplace Intelligence Report, 2020, available at: https://www.oracle.com/explore/ai-at-work-2021/pf-AI-Work-study-1/?source=:ow:o:h:mt:::RC_WWMK200506P00038C0003:EM21_HCM_Q4_C28_M0101_S031YZ18_DS125_T10&intcmp=:ow:o:h:mt:::RC_WWMK200506P00038C0003:EM21_HCM_Q4_C28_M0101_S031YZ18_DS125_T10&lb-mode=overlay.

[32] *Ai @work: un nuovo alleato per l’uomo*, Gaia Flerter, Industrie Quattro Punto Zero, 2020, available at https://www.industriequattropuntozero.it/2020/10/14/aiwork-un-nuovo-alleato-per-luomo/.



Studying the disaggregated data of such kind of surveys[33], it appears that, according to each worker subjective profile, the same situation could have different impacts and hit in a different way: managers will have different needs compared to employees, and among employees the young ones, will respond differently than the old ones. Similarly, workers with different tasks or people of different gender, age, culture, as well as demographic characteristics and social conditions, just to name a few, will "stay" differently in the same situation.

Therefore, the adoption of AI tools to support the psychophysical well-being of employees should vary from case to case and be tailored to their different situations and conditions, or, in other words, it should keep in mind the existing "**inequalities**" hereby intended **as "differences in needs"** and contribute to overcome them.

Even more importantly, though, companies shall be aware of the "dark side" of the concept of inequality, which is **"discrimination"**, and intervene accordingly.

In this respect since 2015 Enel Group is committed to promoting the active involvement of everyone and introduced a "*Diversity & Inclusion Policy*" based on the fundamental principles of non-discrimination, equal opportunities and dignity.

Among others, **disability** is one of the priority areas in Enel's approach to diversity, together with gender, age, nationality and the promotion of a culture of inclusion and in this respect Enel as other multinational company took off the "Valuable 500" initiative.

[33] *There's no one-size-fits-all approach to mental health at work*, *Mental Health at Work Requires Attention, Nuance, and Swift Action*, Oracle e Workplace Intelligence Report, 2020, available at file:///C:/Users/E351252/OneDrive%20-%20EDP/EDP%20luglio%202018/personale/CONTEST%20ENEL/materiali/companies%20papers/hcm-ai-at-work-volume-2.pdf.

The aim of such initiative is to get the potential that can be generated by the inclusion of people with disabilities around the world.

Valuable 500 is a *global movement* that put disability on the business leadership agenda. In May 2021 has been announced that it has been reached the milestone of securing **commitments from 500 global CEOs** and their companies worldwide to be the tipping-point for change and help unlock the social and economic value of people living with disabilities across the world.

In order to contribute to achieve this goal, Enel was a first mover is reinforcing its actions to address the obstacles that people living with disabilities have to face, integrating their perspectives in the overall business approach.

We believe that adopting inclusive hiring practices and using artificial intelligence (AI) tools will mitigate unconscious bias for fair selection of applicants and will rapidly help in reaching the target of hiring teams equally balanced in terms of gender, ages, “dis-abilities and talents” and allowing at the same time the company to benefit a higher level of innovation, an increased customer base and greater productivity.

AI is also used for hiring people and recruit new talents across different industries. As known LinkedIn is a platform that connects job seekers with recruiters and/or employers. To employees, LinkedIn’s AI suggests a job they may be of interest by automatically reading the profile and the job experience and matching it with the profile that employers are looking for making also suggestions and connections to potential similar job vacation worldwide or tailored by location. Technology is also used for the evaluation of job applicant’s speech, tone of voice, facial expressions and body language during video interviews.

Despite its convenience, AI, however, is also capable of being biased if the algorithms are structured and designed to be based on gender, culture, race, dis-ability status etc. and can be used in order to exacerbate employment systemic discrimination. By way of example, assessing just facial movement and voice may discriminate all people with disabilities that affect tone of voice and facial expression such as speech disorders or blindness. Also, tests on personality made on line and used in combination with AI tools may screen out people with illnesses even mental.

In general, AI tools for hiring are programmed to find out ideal profiles based on employer’s preferred candidates which normally are based on the employer’s existing basket of employees’ profile. This means that, if people with disabilities are not represented in the employer’s basket, then the AI hiring tool learn to screen out job profiles with a disability. Essentially, underrepresented profiles are treated by AI tools as undesired profiles. As a result, people with disabilities risk being excluded. To overcome bias, AI hiring tools need to be designed and trained with diverse data including the ones of employees with disabilities.

Considering that as of today disabled people are underrepresented in the workforce if AI tools are not well designed and trained this phenomenon will be, unsurprisingly, emulated with tremendous impact in terms of inequalities.

The danger therefore of AI, and algorithm's that lacks of "diversity" data, is that may enlarge existing patterns of exclusion harming deeply communities that are already treated as "diverse". This is also true in case of gender inequality and stereotypes. Algorithms have the potential of spreading and reinforcing such harmful stereotypes, with the effect of further marginalizing women on a global scale.

If used in the proper manner conversely AI has the potential of being part of the solution and pushing gender equality in our societies.

But what does it mean "proper manner"? It simply means that AI is to be used in order to fight against inequalities and filling all those gaps that "humans" are unable to fight against due to historical and cultural stereotypes.

As known SDG 5 covers a diverse range of targets that aim to eliminate inequalities especially for gender and drives constructive work across different areas promoting the use of technology, to empower women. By leveraging more on results, and not on *time-in-seat*, with an intensive use of remote work and flexible schedules gender equality will be more easily reached at least at work.

Has been demonstrated that parents who are able to flex their schedule and work when it is most convenient for them can often maintain a professional job and perform at a high level. Flexibility is key to reduce or even eliminate barriers to employment and may help people to find the proper job the job that meet the needs of the employee. Such sustainable approach will turn into value for the Company and the Employer.

Due to (or thanks to) Covid-19 pandemic, we have all learned that remote working is feasible and can better accommodate needs of people who have historically been left out of the traditional workforce.

Whether it is due to medical issues, mental health issues, a disability, a rural location without many job opportunities, or lack of access to transportation, many people are unable to find a job that allow to live in a decent manner.

Gender inequalities with the Covid-19 pandemic are amplified, and the areas where we were registering some improvements are unfortunately regressing. As known women are still used to carry out the main burden of unpaid work, and now with the Pandemic this tendency will grow. For instance, a poll covering seventeen countries on unpaid work reveals that women and girls have taken over a large share of responsibility regarding childcare, family, and household duties due to the Covid-19 pandemic. As commonly known unpaid work has big implications in terms of grow, income, career opportunities, and in general but more important on women's mental well-being.

AI will fill the gap and the real goal is not only about achieving gender balance but by analyzing the context of both sexes via a gender analysis, ensuring that all voices equally and fairly are heard, and creating an equal playing field.

Having recalled some of the main inequalities at work, in the following paragraph we present some AI tools offering valuable solutions to eliminate some of the described gaps and making life easier and happier for the lucky employees that will enjoy them and allow organization to achieve stronger job performance, job satisfaction, and commitment.

3.2 Improvements in terms of social sustainability

AI can be used as a booster to improve the quality of life of people, namely – in this paper – of employees, while overcoming diversities and helping disadvantaged individuals finding a better work and life condition.

The systematic implementation, application and use of what we have called "AI for the public good" has a big impact in terms of "social sustainability" with particular focus on the secondary and indirect benefits that can generate on employees, but also on other categories of subjects, as well as on the positive rewards for the companies adopting such virtuous behavior and for their business.

Before digging into the core this final topic, it is worth spending few words on the concept of "social sustainability" itself, giving a definition and briefly illustrating its importance.

Social sustainability is one of the three pillars of "sustainability" – alongside "environmental sustainability" and "economic sustainability". Among the many definitions available, in this paper we will focus on the following: "***Social sustainability is about identifying and managing business impacts, both positive and negative, on people***".

In a balanced system, the three souls of sustainability "Planet, People, Profit" should be equally valued: instead exploring the topic it appears that social sustainability is largely neglected in mainstream sustainability debates. In fact, traditionally, attention is given in the first place to the economic aspects of sustainability; secondly, but only over the last decade, after the environmental movement took shape and gradually grew in importance, the environmental component of sustainability has risen to priority too; while the interest in social sustainability remains low and its importance underestimated.

Social sustainability is a multidimensional concept whose core substance and added value is only fully revealed in its interrelationships with other sustainable development pillars and dimensions, and it is perhaps because of this complex nature that it struggles to emerge.

Social sustainability has five dimensions: **equity, diversity, social cohesion, quality of life, democracy-governance**. All of them should be screened to evaluate whether a business or a project, in the case at stake, whether the implemented AI tool, is socially sustainable

Below some useful questions to guide such assessment could be:

1. Equity

Will it assist the target group to have more control over their lives, socially and economically?

Will it identify the causes of disadvantage and inequality and look for ways to reduce them?

2. Diversity

Will it allow for diverse viewpoints, beliefs and values to be taken into consideration?

3. Social cohesion

Will it result in the provision of increased support to the target group by the broader community?

4. Quality of life

Will it improve mental health outcomes for the target group?

Will it improve the safety and security for the target group?

5. Democracy and governance

Will it have a budget sufficient to ensure adequate delivery by qualified trained staff?

Will it ensure that the use of volunteers is appropriate and properly governed?

A complete template of the assessment is available under **Appendix 2**.

Numerous companies have already adopted the described principles following them as guidelines for their technological development.

Tech companies, for example, for each AI application introduced and installed, requires the following features to be ensured: 1. Being socially beneficial, 2. Avoid creating or reinforcing unfair bias, 3. Be built and tested for safety, 4. Be accountable to people, 5. Incorporate privacy design principles, 6 Uphold high standards of scientific excellence, 7. Be made available for uses that according with these principles.

4. AI to overcome inequalities in the psychophysical well-being field – practical applications

In order to pursue the SGD3 in compliance with the legal framework and characteristics, developers must consider at least the following issues when they are releasing new algorithm.

4.1 Application: Tool for Next New Normal

a. **Context**

Progress in vaccination campaigns in some countries suggests the end of the emergency phase and, consequently, the possibility of return to the office. Nonetheless most people do not believe workers will return to the office full-time after the coronavirus pandemic while seems that a hybrid method of working, partially in presence, for activities such as team-buildings, staff meetings, on boarding etc. and often remotely. The return to the office will not be the same for everyone: each employee will have different needs, changing over time, and the purpose will be to find the optimum balance while guaranteeing the well-being of all the stakeholders and in particular of all employees of big companies and multinational worldwide. Remote working gives people more time and energy to focus on their dietary program, relationships, leisure, and passions—all contributors to better well-being and healthier lives. It can improve the mood and the happiness at work and at home. In compliance with the SDG 3, the employer should try to accommodate these needs and at the same time improving achievement of company objectives. Below we propose the application of artificial intelligence to achieve this ambitious purpose.

b. **Proposal**

Provide employees with a web platform, responsive device, which organizes workstations, meeting rooms, relax zone, gym space etc. in a dynamic and flexible approach according to the daily needs of employees.

c. **Needs & Constrains**

This platform will have to proactively organize the spaces on the basis of the numbers of the people of a team and if the space in the open space allows it, on the basis of the data of each employee (for example from e-profile), create a place where each employee may find colleagues with **similar interests and passions** (both personal and professional) on one side and people (in smaller numbers) who give a different vision / interests on the other (**antagonist interests and passions**). More and more often, artificial intelligence, through the profiling of cookies, tends to flatten our imagination by offering us articles / experiences "similar" to those we have previously searched for. Track records of antagonist behavior are crucial for the success of no discrimination. The inclusion of people "out of context" is aimed precisely at breaking down this trend and increasing the culture of diversity.

The second needs is to allow everyone to use the company spaces in the same manner; one of the dangers employees of different companies worldwide encountered in the past is that of having a number of people book themselves systematically faster than others, thus creating an obvious

disparity. To overcome this phenomenon, we believe that there are simple measures that can be introduced easily in most of the companies with a significant number of employees and spaces and workstations that have – especially after the pandemic – new shapes.
To be practical and pragmatic here we propose the following measures:

1. Associate each employee with a ranking based on the number of hours / days spent in the office per month (this ranking is reset every month, to meet particular business needs) and on the reliability of the booking (therefore all the colleagues who book and then do not go respond to the desk reservation request in the office by time slots, so for example, the basic setting (i.e. the initial configuration), the time slots and their availability are
   00:00 - 10:30 - It is possible to reach 40% of the total availability
   10:30 - 13:00 - It is possible to reach 70% of the total availability
   14:00 - 18:30 - It is possible to reach 90% of the total availability
   18:30 - 00:00 - It is possible to reach total availability

   The system, once each time slot has been completed, will sort the subscribers on the basis of the ranking, therefore for example for the first time slot up to 40% of the total available capacity, for people who do not fall within that time slot it will propose through geo-location other locations available within a radius of 50 km. The artificial intelligence based on the habits of users on a weekly basis (every Friday) will evaluate the division of time slots and the total availability of spaces. If at 18:30, 90% of the total capacity has not yet been reached, the system will fish by ranking order all the excluded employees.

2. Respect the needs of entire teams as well as individuals, the organization of several people within a team is more complicated, the system must therefore guarantee a method for registering spaces on a daily, weekly and monthly basis. In fact, we can find the young employee who needs to breathe as much as possible the company reality to get to know it and create his own network, the parent who has to organize the custody of the children or the manager of a team that has to brainstorm. With the aim of not creating inequalities, the system must allow everyone to use the company spaces by creating "buffers" of workstations to allow the management of all needs.

3. We also think that psychophysical well-being also passes through the stimulation of new ideas and we believe that the value of each of us can also increase through the exchange of experiences and diversity. During the COVID period, companies have continued to hire and other colleagues have changed units, our platform in order to increase contamination can proactively propose through pop-ups, based on some criteria (including availability on the agenda), meetings (lunch or simply a coffee break) in order not to create the roles of those who request and those who accept / reject.

4.2 Application: Tool for Right to Disconnect

a. **Context**

During long periods of smart working, one of the advantages is that the employee, always in compliance with company policies / objectives, can manage time more independently than in the past. This great opportunity, however, can present risks for some groups, in fact a not very clear delimitation between private life and working life, can feed an unhealthy work-life balance.

b. **Proposal**

Employee health and wellbeing is also affected by a proper work life balance. Even a few minutes break done in a systematic way can improve the quality of life! With the new ways of working there is the risk of having, for example, consecutive meetings in which the only real break is the lunch time. For example, companies may introduce a monitoring system that forces the employee – that is permanently stuck in meetings - to have a break during its working day.

c. **Needs & Constrains**

Monitoring during working hours is a very delicate issue, over time other companies have tried through digital tools to favor the work-life balance (for example by switching off the mail servers at certain times). We think this can be a very strong action and in order to be eventually developed, an ad-hoc study and major change management are required. Based on the user experience practiced during the evacuation tests, we propose that after a certain number of consecutive hours the computer goes into lockdown for a few minutes. However, artificial intelligence will have to completely autonomously and without the possibility of tracing staff data to understand if the employee needs to take a short break. The data that could be used are therefore for example the movement of the mouse, the presence in calls and other inputs.

5. Conclusions

The application of Artificial Intelligence in the daily life to fill the gaps of inequalities is not only in line with the UN SDGs, but also with the UN Global compact principles, according to which businesses should i) support and respect the protection of human rights, and ii) make sure that no human rights’ abuses are carried out. The benefits to companies, employees, and the environment are clear. Mental and phisical well-being must be a priority for Employers.

Multinational companies can contribute proactively to be part of this challenge.

AI is the tool enabling the overcome inequalities, protect health and wellbeing with the aim of creating shared value for all stakeholders.

The role of each employee is crucial for the achievement of this ambitious goal.

APPENDIX 1: Scheme for Development of AI system

Usually, the development of an AI system (in this case of Machine Learning) takes place in compliance with an ordered sequence of phases which, for the sake of simplicity, we can summarize as follows.

Analysis of the problem

The first step in developing the system is to identify the object of the forecast, i.e. what you want to predict. This "response" is usually represented by a variable, called target. The imaginable scenarios are innumerable. By way of example, we could hypothesize a need consisting in forecasting the sale price of a property, or we could hypothesize the need to classify a review as positive or negative. In both cases the answer to the question that arises will always consist of a number, but with different meanings.

In the case of the forecast of the sale price of a property, in fact, the answer to the question will be a real number such as 100,000 euros or 235,400.50 euros. In the case of the review classification, on the other hand, there will be a 0 (positive) and a 1 (negative or not positive) as possible answers.

This is an absolutely essential phase for the correct development of an AI system, as it is essential to understand, for example, whether it is a regression or classification problem, in order to identify the most suitable algorithms for the use case.

Choice / formation of the dataset

Having precisely identified the problem to be solved, it is necessary to identify the "data" to be used for learning. It is worth remembering that AI systems have as an unfailing premise the use of large amounts of data for machine training. Returning to the previous examples, the dataset can consist of statistics on the sales of real estate units made in a specific city in a defined period or in all the reviews published on a certain product.

As regards the dataset, the main need is to have quality data available and suitable for the chosen learning model. In supervised training it will be, for example, having correctly labeled data. This results in the need to have for each element of the whole, the indication of “n” characteristics (features) that may be relevant for the prediction to be made and, above all, the target variable. Trivially, for the prediction of the sale price of a property, characteristics such as: the size, the plan, the position, etc. may probably be useful. The target variable, on the other hand, will be the actual sale price for that particular property.

The choice / formation of the dataset is an absolutely crucial phase, especially with a view to avoiding or minimizing the risk that the application of AI may cause discrimination against individuals.

Model training

The goal of applying AI technologies is to make predictions on unknown data.

To achieve this result, the machine must be able to "learn" from the data in its possession. This is done by applying a learning algorithm to the dataset and at the end of the training phase the model to be used for future predictions will finally be generated.

The choice of the algorithm, as we have already highlighted above, depends on the type of problem to be solved.

For example, if we wanted to classify a tumor as benign or malignant depending on the graphical representation, we could use a logistic regression algorithm. The goal of this algorithm is to predict the probability that a tumor belongs to the class of malignant (1) or non-malignant (0). In the learning phase, the machine will try to learn the function (or more simply that mathematical relationship between the data) that minimizes the possibility of error[34], using specific coefficients (ie the "weights") for this purpose.

Evaluation of the accuracy of the model and improvement

Once the algorithm has been trained, it is necessary to test it to verify its operation.

An additional dataset will therefore be required to carry out the aforementioned checks. In practice, the initial dataset is divided into two subsets: the first, larger (about 70% of the data) to be allocated to the actual learning phase; the second, more contained (about 30% of the data) to perform the tests.

At this point, the developer will be able, in different ways, to verify the goodness of the generated model. In the presence of a supervised learning model, it will be easy to carry out these tests, comparing the prediction made by the system with the correct output (present in the dataset).

Using the model to make predictions

Once the verification phase is complete, if no adjustments are required to the model, it can actually be released for actual use.

If, on the other hand, anomalies (e.g. poor accuracy) were found at the outcome of the verification, it will be possible, instead, to proceed with a new training, making the appropriate precautions.

As already highlighted, the activities to be carried out in practice can be influenced by the type of algorithm used: it is evident that the development of a model with unsupervised learning will be significantly different from supervised learning, if only for the fact that the data starting point will not be labeled and, therefore, the machine itself will have to extrapolate the relevant characteristics from the data.

[34] BROWNLEE J., *Mastering Machine Learning Algorithms.*

APPENDIX 2: Template questions for the assessment

Equity

Will the project [AI tool] reduce disadvantage for the target group?

Will it assist the target group to have more control over their lives, socially and economically?

Will it identify the causes of disadvantage and inequality and look for ways to reduce them?

Will it identify and aim to meet the needs of any particularly disadvantaged and marginalized people within the target group?

Will it be delivered without bias and promote fairness?

Diversity

Will the project [AI tool] identify diverse groups within the target group and look at ways to meet their particular needs?

Will it recognize diversity within cultural, ethnic and racial groups?

Will it allow for diverse viewpoints, beliefs and values to be taken into consideration?

Will it promote understanding and acceptance within the broader community of diverse backgrounds, cultures and life circumstances?

Social cohesion

Will the project [AI tool] help the target group develop a sense of belonging in the broader community?

Will it increase participation in social activities by individuals in the target group?

Will it improve the target groups' understanding of and access to public and civic institutions?

Will it build links between the target group and other groups in the broader community?

Will it result in the provision of increased support to the target group by the broader community?

Will it encourage the target group to contribute towards the community or provide support for others?

Quality of life

Will the project [AI tool] improve affordable and appropriate housing opportunities for the target group?

Will it improve physical health outcomes for the target group?

Will it improve mental health outcomes for the target group?

Will it improve education, training and skill development opportunities for the target group?

Will Social Sustainability Improve the Employment Opportunities for the Target Group?

Will it improve access to transport for the target group?

Will it improve the ability of the target group to meet their basic needs?

Will it improve the safety and security for the target group?

Will it improve access to community amenities and facilities for the target group?

Democracy and governance

Will the project allow for a diverse range of people (especially the target group) to participate and be represented in decision-making processes?

Will the processes of decision-making for the project be clear to and easily understood by staff and stakeholders?

Will it have a budget sufficient to ensure adequate delivery by qualified trained staff?

Will it ensure that the use of volunteers is appropriate and properly governed?

Will the duration of the project be sufficient to achieve the desired outcomes?

Have you considered what will happen when the project [AI tool] ceases [to be used/applied]?