Unlocking the Synergy: Artificial Intelligence and (old and new) Human Rights

Carlo Casonato*

ABSTRACT: Artificial intelligence (AI) deeply and pervasively impacts our lives. In this short paper, I propose two lines of thoughts aimed at updating the catalog of human rights in light of the potential and risks of AI (mainly Machine Learning). The first considers the adaptation of certain traditional principles (informed consent and non-discrimination) to the challenges posed by AI. The second covers four new rights, built upon the specific characteristics of AI systems, with the aim of effectively addressing AI pros and cons.

KEYWORDS: Artificial intelligence; AI act; fundamental rights; informed consent; non-discrimination

1. Introduction

In this short paper, I would like to propose some thoughts on how both traditional and new generation rights can provide a framework within which new AI technologies may operate securely and in service of individuals and society. In this regard, instead of being seen as obstacles and impediments to scientific progress, law and rights can synergize with new technologies to make AI Trustworthy and Human-centered.

Specifically, I will address three main points.[1]

Firstly, I will argue that AI is not just another technology. Its distinctive characteristics (such as autonomy, unpredictability, lack of transparency) and the profound impact it wields over society and our daily lives distinguish it as a technology in a league of its own. Essentially, AI is characterized by an unprecedented pervasiveness and transformative potential.

* *Professor of Comparative Constitutional Law, Law School, University of Trento; Jean Monnet Chair on AI and EU Law (T4F); Editor-in-Chief of the BioLaw Journal. Mail: carlo.casonato@unitn.it. Invited contribution.*

[1] The paper revisits the speech given at the international seminar organized by the European Public Law Organization on 15 September 2023. Some of the presented results are part of the activities of the NextGenerationEU project (FAIR – Future AI Research – PE000013) co-funded by the European Union. The paper maintains the discursive style of the presentation, limiting bibliographical references to the essential. The views and opinions expressed are solely those of the author and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

Confronted with these distinctive features, my second point will explore how certain old, established rights can be adapted to grapple with the novel challenges posed by AI. I will refer to the principles of informed consent and non-discrimination.
Lastly, in the third point, I will outline a series of four 'emerging' rights that I believe warrant careful consideration. These rights play a crucial role in guiding the development of AI towards enhancing the well-being of individuals and advancing society, both in the present and the future, while upholding the well-established principles of Trustworthy and Human-centered AI.

2. The characteristics of AI

The extensive reach of AI, its profound and pervasive influence on our lives, enables it to reshape the significance we attribute to our reality and experiences. In this context, some authors refer to the re-ontologizing capacity of AI, suggesting that AI is indeed reshaping the world. Thus, the concepts of 'onlife' and 'infosphere' are used to state that we, especially the younger generations, now live in a world where AI is present everywhere and significantly influences many of our daily activities.[2] Furthermore, for AI to achieve its full potential and optimize its performance, it necessitates a conducive context and an environment built around it. For instance, we are reshaping our cities into smart urban centers, employing sensors and cameras capable of capturing, transmitting, and processing extensive amounts of data pertaining to our lives. In the same way, we are modifying our homes to make smart home automation increasingly efficient. In essence, we are adapting the we inhabit (and perhaps ourselves) to be 'AI-friendly'.

Some authors talk about an "envelope" that we are constructing around AI, and ultimately around ourselves.[3] This trend encapsulates the transformative capacity of AI. However, what's crucial is that this envelope is built in a way that respects the human dimension, safeguarding aspects such as our privacy and the protection of personal data. It is imperative that the pervasive and transformative potential of AI serves the majority of humanity, rather than becoming a tool for exploitation by a few to the detriment of many.[4]
In this sense, the concept of 'digital constitutionalism' can be invoked, whereby alongside the necessary limitation of powers, it is imperative to reconsider the list of human rights (understood this time as those of the human being in the face of technology), in order to ensure that AI becomes a tool for democracy, social and economic progress, and not a new vehicle for inequalities and discriminations.[5]
So, which rights should we invoke to meet the new needs for protection and promotion of the human being in the context of AI?

[2] L. Floridi, *Etica dell'intelligenza artificiale*, 2022, 53 ss.
[3] L. Floridi, *Etica dell'intelligenza artificiale*, cit., 56 ss.
[4] S. Zuboff, *The Age of Surveillance Capitalism. The fight for a human future at the new frontier of power*, 2019; C. O'Neil, *Weapons of math destruction: how big data increases inequality and threatens democracy*, 2016; K. Crawford, *Atlas of AI, Power, Politics, and the Planetary Costs of Artificial Intelligence*, 2021; M.R. Ferrarese, *Poteri Nuovi*, 2022.
[5] See, in general, G. De Gregorio, *Digital Constitutionalism in Europe*, 2022; L. Torchia, *Lo Stato digitale*, 2023.

ISSN 2284-4503

3. Old rights and principles

3.1. Informed consent

As mentioned earlier, certain traditional rights can be adapted to the new challenges posed by AI. One initial old right that can be adapted is informed consent. We know that some AI systems have advanced to such a degree that they can be mistaken for humans. Engaging in dialogue with ChatGPT, for instance, gives a distinct sensation of conversing with one of us, a human being, or a human-like entity that might be even more intelligent and intuitive than ourselves; yet it is equally evident that the same application occasionally provides responses that are completely inaccurate or entirely fabricated.[6]

Faced with AI systems that we can no longer distinguish from humans (think of the Turing test and its Imitation Game),[7] it is important to recognize the right to know whether we are interacting with a person or a machine: whether my interlocutor is a peer or an AI system. It is indeed crucial to avoid the confusion that might otherwise arise and to be clear about the nature of our interlocutor. This way, misunderstandings or breaches of trust due to misplaced expectations of empathy and confidence can be avoided, such as trusting doctor-robots, official-robots, or even companions that we believed to be human. This right, which could be viewed as a new version of informed consent, is already acknowledged in the European proposal for AI regulation (AI Act), where Article 52 stipulates that "Providers shall ensure that AI systems intended to interact with natural persons are designed and developed in such a way that the AI system, the provider itself or the user informs the natural person exposed to an AI system that they are interacting with an AI system in a timely, clear and intelligible manner (…)."[8] Additionally, at a national level, France amended the "*Loi Bioéthique*" in August 2021, stating that a doctor who decides to use AI for their profession "must ensure that the person concerned has been informed and is... informed of the interpretation that results from it".[9]

In addition to information about the nature of our interlocutor, it is also important to have insights into the reasons that led the machine to generate a certain outcome. This aspect is complicated by the fact that the most advanced AI systems (machine learning, deep learning, neural networks) operate with non-transparent, opaque internals. Given their complexity (for example, ChatGPT utilizes 175 billion different parameters), even programmers are unable to understand how the system generated the output. This is the phenomenon known as the 'black box', which, by concealing the inner workings of the system, presents two primary challenges.

[6] For a recent case, see B. Weiser, *International New York Times*, May 30, 2023: *A lawyer used A.I. to write a cour filing. It backfired.*

[7] A. Turing, *Computing Machinery and Intelligence*, in *Mind*, 1950, 433.

[8] See the Amendments adopted by the European Parliament on 14 June 2023 on the Proposal for a Regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (*Artificial Intelligence Act*) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD).

[9] Article L. 4001-3 of the "*Code de la Santé Publique*". See C. Crichton, *L'intelligence artificielle dans la révision de la loi bioéthique*, in *Dalloz Actualité*, 16 settembre 2021.

The first issue pertains to the legitimization of functions performed with AI. How can a doctor, for example, present a diagnosis without being able to explain the underlying reasons? How can a public official refuse or grant a license without providing the rationale behind their decision? Similarly, how can a judge issue a decision with evident deficiencies in the reasoning? In essence, the black box poses a risk to the acknowledgment and legitimacy of activities that employ AI systems.

The second problem concerns rectifying errors that the system might make, given the extreme difficulty, and sometimes impossibility, of pinpointing where the machine went wrong.

In this regard as well, the European Union's proposed AI regulation introduces a right that can be seen as a new iteration of the principle of informed consent. Article 13 of the amended version, in particular, stipulates that "High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable providers and users to reasonably understand the system's functioning."[10]

3.2. Non-discrimination

A second established principle that can be adapted for the use of AI is non-discrimination. It is well-known that the outputs generated by AI can be biased, erroneous, imprecise, and lead to discriminatory effects. Numerous studies have already underscored this aspect, particularly within the realm of justice. For example, algorithms are utilized by judges to evaluate various factors, including the social risk presented by an arrested individual (as seen in the COMPAS case). Similarly, in medicine, some research has pointed out the risk that AI utilization could exacerbate existing discriminations in certain healthcare services (like the US healthcare system), potentially leading to a form of "race-based medicine".[11]

Experts are actively working to address the discriminatory effects of AI, yet the problem remains unresolved. 'Cleaning' the datasets used, by incorporating accurate and comprehensive information, is a necessary step toward achieving accurate outputs, but it does not completely solve the problem. Even with accurate information, the statistical-probabilistic logic that AI operates on can arrive at conclusions that discriminate, for instance, based on ethnicity, race, or gender[12]. This occurred, for example, when a justice-related AI system inferred that since the percentage of African Americans in

[10] The amendment continues: "Appropriate transparency shall be ensured in accordance with the intended purpose of the AI system, with a view to achieving compliance with the relevant obligations of the provider and user set out in Chapter 3 of this Title.
Transparency shall thereby mean that, at the time the high-risk AI system is placed on the market, all technical means available in accordance with the generally acknowledged state of art are used to ensure that the AI system's output is interpretable by the provider and the user. The user shall be enabled to understand and use the AI system appropriately by generally knowing how the AI system works and what data it processes, allowing the user to explain the decisions taken by the AI system to the affected person pursuant to Article 68(c)".

[11] D.A. Vyas et al., *Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms*, in *The New England Journal of Medicine*, 2020, 874-882; A. Bracic, et al., *Exclusion cycles: Reinforcing disparities in medicine*, in *Science*, 2022, 6611, 1158-1160

[12] Among others, F. Zuiderveen Borgesius, *Discrimination, artificial intelligence, and algorithmic decision-making*, Directorate General of Democracy, Council of Europe, 2018.

US prisons was higher than that of Caucasians, race was an indicator of dangerousness. From a statistically real fact, the machine drew an obviously incorrect and highly discriminatory correlation.
Resolving this issue is not easy. One approach is to refine the 'cleaning' of datasets, ensuring correct, complete, and up-to-date data. The European Union's proposed regulation, as amended by the EU Parliament, aligns with this direction, recommending that datasets used for high-risk AI systems shall be "relevant, sufficiently representative, appropriately vetted for errors and be as complete as possible in view of the intended purpose. They shall have the appropriate statistical properties, including, where applicable, as regards the persons or groups of persons in relation to whom the high-risk AI system is intended to be used" (Article 10.3). Nevertheless, the problem persists, as the correctness of initial data does not guarantee that the results will always be accurate.
To address this challenge, tackling the first of the 'new rights' can be useful: the right to "Human in the Loop". That is, the right to decisions that are not solely made by an artificial system. In essence, it entails the right to be recipients of decisions subject to significant human oversight, to mitigate potential errors and discriminatory outcomes from machines.

4. New rights

4.1. Human in the Loop and the right to the hero

It is well known that machine learning systems can operate with broad margins of autonomy and unpredictability. This is one of the reasons why they are used. Due to this property, and in the face of the advantages as well as the risks indicated, both ethics and law have invoked the principle of "Human in the Loop". Specifically, it is recommended that decisions made with the assistance of AI are always supervised and controlled by a person who takes responsibility for the decision itself. Considering that a machine can decide without transparency (the phenomenon of the black box), make mistakes, or reach discriminatory outcomes, it is indeed essential that a human being intervenes to oversee and potentially correct the result.
In Europe, this right is already partially recognized under the General Data Protection Regulation, in Art. 22 which states "the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her". The AI act confirms and reinforce this approach, underscoring the importance of the human oversight. Art. 14.1 of the mentioned amended version provides that "High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they be effectively overseen by natural persons as proportionate to the risks associated with those systems."
Despite this commendable effort, however, there remains a concrete risk that the principle of human oversight may remain merely a formal and fictitious element. How many public administration officials, for instance, possess the computer skills necessary to understand and accurately interpret the outputs of a system whose inner workings even their designers are ignorant of? And how many will take on the burden of justifying a decision that deviates from what the machine dictates? In an age of widespread shirking of responsibility, furthermore, who will take the personal risk of defying an output generally perceived as correct? The risk, as Antoine Garapon puts it, is the "effet moutonnier"

(herd or sheep effect): a phenomenon of standardization and de-responsibilization stemming from the fact that the decision is actually "captured" by the algorithm, rendering the principle of human oversight merely superficial, proclaimed but not effectively and practically applied.[13] To exercise it, in fact, would require a professional who not only possesses specific computer science expertise but is also strongly motivated to counter what they deem incorrect, personally assuming the responsibility and corresponding risk; it would require – one might say – a true hero. In this sense, paradoxically, one could speak in terms of a 'right to the hero' or a 'right to heroism'.

In the face of this risk, the version approved by the European Parliament of the AI Act intervenes, attempting to reinforce and, to some extent, safeguard the position of the human being overseeing the system. Article 14.1 stipulates that "Natural persons in charge of ensuring human oversight shall have sufficient level of AI literacy in accordance with Article 4b and the necessary support and authority to exercise that function, during the period in which the AI system is in use and to allow for thorough investigation after an incident".[14]

4.2. The right to discontinuity

The second of the new rights that could be here proposed is connected to the statistical-probabilistic approach with which AI operates. In particular, the profiling to which all of us are subjected is based on what we could call our 'historical self', consisting of preferences, orientations, and decisions as we have expressed them in the past. An example of this is when platforms on which we book vacations, order meals, or choose movies suggest options that correspond to what we have booked, ordered, and chosen up to that point. The risk, therefore, is to become trapped in a past that is impervious to potential new interests, curiosities, and changes. The problem intensifies when profiling is used, for instance, in the realm of political orientations, where young people, particularly, are 'bombed' with information perceived as comprehensive and objective but that has actually been selected for them, in order to cater to their preferences and choices. The result of this profiling is placing them within a comfort zone, an echo chamber whose effect, in the absence of true confrontation, is a progressive and radical polarization of their own ideas.

[13] A. GARAPON, J. LASSÈGUE, *Justice digitale. Révolution graphique et rupture anthropologique*, Paris, 2018. See also, among others, A. SIMONCINI, *L'algoritmo incostituzionale: l'intelligenza artificiale e il futuro delle libertà*, in *BioLaw Journal – Rivista di BioDiritto*, 1, 2019, 63-89.

[14] Art. 4b establishes the content and requirements for AI literacy: "1. When implementing this Regulation, the Union and the Member States shall promote measures for the development of a sufficient level of AI literacy, across sectors and taking into account the different needs of groups of providers, deployers and affected persons concerned, including through education and training, skilling and reskilling programmes and while ensuring proper gender and age balance, in view of allowing a democratic control of AI systems. 2. Providers and deployers of AI systems shall take measures to ensure a sufficient level of AI literacy of their staff and other persons dealing with the operation and use of AI systems on their behalf, taking into account their technical knowledge, experience, education and training and the context the AI systems are to be used in, and considering the persons or groups of persons on which the AI systems are to be used. 3. Such literacy measures shall consist, in particular, of the teaching of basic notions and skills about AI systems and their functioning, including the different types of products and uses, their risks and benefits. 4. A sufficient level of AI literacy is one that contributes, as necessary, to the ability of providers and deployers to ensure compliance and enforcement of this Regulation."

The important issue here is that of a profiling entirely focused on the past, a 'conservative profiling' that effectively leads to being imprisoned in a virtual bubble that can easily be mistaken for the real world and prevents genuine engagement with the new. From this perspective, to preserve a minimum of curiosity, doubt, and a desire for change (or we could say, authentic freedom), one could invoke a kind of new right to discontinuity or inconsistency; a right to "step out of the bubble", to abandon our 'past self' in order to be stimulated to engage with the different, to embrace the richness of contradictions and unexpected novelties.

4.3. The right to a human environment

A third new right is linked to the mentioned pervasive and transformative scope of AI. To optimize the functioning of the most advanced systems, it is necessary to build around them an environment suited to make them operate most efficiently (the mentioned Floridi's envelope phenomenon). In some cases, as in new large airports, the envelope has already been constructed to allow AI, which guides the planes rather than human pilots, to function optimally. This trend of constructing "AI-friendly" environments is increasingly spreading, resulting in what could be described as a Midas touch effect, where things that come into contact with technology are transformed[15]. In this regard, attention must be paid to ensure that this type of locations does not produce the effect of excluding humans from the environments they have always lived in, preventing them from recognizing their own city due to the expansion of features of a pervasive smart city, or their own home due to growing home automation. If this were to happen, the envelope could turn against humanity: somewhat like what happened to King Midas, who starved to death because everything he touched turned into gold, ending up in solitude. If this were to happen, we would risk living in a world tailored to AI, rather than tailored to humans. For this reason, the envelope must be limited to environments where we want to prioritize the performance and effectiveness of AI. And a right, the third new right on this list, can be invoked: the right to live in an environment that remains human-centered.

4.4. The right to AI

So far, AI has been seen as an expression of power that can jeopardize our interests as human beings. This is the phenomenon whereby private powers come to perform essential functions, functions of a public nature, such as education, healthcare, and justice. From this perspective, a variety of the new digital constitutionalism imposes, paraphrasing the 1789 *Déclaration des droits de l'homme et du citoyen*, the limitation of powers, including those of private entities that dominate the AI world, and the guarantee of rights. For this reason, a list of rights (both old and new) has been so far proposed in some way "against" AI.

However, AI also offers enormous potential to enhance our existence. From medical diagnostics to urban traffic control, from agriculture to clinical research and the fight against the climate change, AI can assist the human professional by performing a multitude of operations more quickly and accurately for the benefit of individuals and society. In these cases, the advantages of employing AI are so high and the risks so reduced that one could advocate for a true right to AI: a right to be recipients of

[15] The concept is distinct from Stuart Russell's King Midas effect concerning values misalignment. See *Human Compatible: Artificial Intelligence and the Problem of Control*, 2019.

more efficient and faster activities and services thanks to the intervention of AI alongside humans. The mentioned Article 22 of the GDPR, thus, could be inverted to ensure “the right to be subject to a decision based on automated processing, in conjunction with human involvement”.
This position could be linked to what was already established in 1948 by the Universal Declaration of Human Rights, which states that "Everyone has the right... to share in scientific advancement and its benefits" (Article 27). In this way, while alleviating the aforementioned risks, the law could serve as a catalyst for technological and social advancement, advocating for AI in areas where it outperforms human capabilities.

5. Concluding remarks

Building upon the points briefly mentioned, there arises a need to embark on a process that aligns legal frameworks and rights with a fast-evolving reality. In this perspective, a viable approach involves reinterpreting established rights and formulating entirely novel ones, all with the purpose of delineating a trajectory toward harmonizing the legal and technological realms. This synergy is intended to cultivate an AI utilization that truly embodies trustworthiness and a human-centered focus. Nonetheless, relying solely on legal measures might prove insufficient in holistically managing AI, effectively mitigating its risks, and harnessing its potential. This underscores the necessity for a more integrative approach, encompassing endeavours such as proposing avenues for raising awareness and initiating tailored educational pathways to cultivate a full understanding of the advantages and pitfalls linked to AI[16]. Given AI's pervasive influence, actions spanning various facets must adopt a profoundly interdisciplinary stance, intertwining the scientific and technological realm with dimensions encompassing politics, ethics, sociology, anthropology, and law.

[16] See L. Floridi, F. Cabitza, *Intelligenza artificiale. L’uso delle nuove macchine*, 2021, among others.