AI & SOCIETY
https://doi.org/10.1007/s00146-022-01453-8

ORIGINAL ARTICLE

A principle-based approach to AI: the case for European Union and Italy

Francesco Corea[1] · Fabio Fossa[2] · Andrea Loreggia[3] · Stefano Quintarelli[4] · Salvatore Sapienza[5]

Received: 29 July 2021 / Accepted: 13 April 2022

Abstract
As Artificial Intelligence (AI) becomes more and more pervasive in our everyday life, new questions arise about its ethical and social impacts. Such issues concern all stakeholders involved in or committed to the design, implementation, deployment, and use of the technology. The present document addresses these preoccupations by introducing and discussing a set of practical obligations and recommendations for the development of applications and systems based on AI techniques. With this work we hope to contribute to spreading awareness on the many social challenges posed by AI and encouraging the establishment of good practices throughout the relevant social areas. As points of novelty, the paper elaborates on an integrated view that combines both human rights and ethical concepts to reap the benefits of the two approaches. Moreover, it proposes innovative recommendations, such as those on redress and governance, which add further insight to the debate. Finally, it incorporates a specific focus on the Italian Constitution, thus offering an example of how core legislations of Member States might contribute to further specify and enrich the EU normative framework on AI.

Keywords AI Ethics · Principles · Values · Recommendations · AI Governance

1 Introduction

Nowadays, Artificial Intelligence (AI) solutions are very commonly employed in many applications and systems to get good results in the execution of specific tasks. From computer vision (e.g., Deng et al. 2009) to voice recognition (e.g., Tsiao et al. 2007), from medical support (e.g., Yan et al. 2006) to elderly people assistance (e.g., Broekens et al. 2009), artificial agents that learn from the context and improve their performance offer new fascinating opportunities.

This, however, cannot but also raise new questions about the ethical and social impacts of AI (Floridi and Sanders 2004). These concerns are not connected only with the technology itself, but also with all the stakeholders involved in or committed to its design, implementation, deployment, and use (Floridi et al. 2018). The aim of this document is to address such concerns to pave the way towards the development of trustworthy AI systems (Quintarelli et al. 2019).

In what follows, we propose a general normative framework and a set of obligations and practical recommendations for the development and use of AI systems. As cornerstones of the approach we assume the ideas and ethical standards that are enshrined in the founding documents of our social organization. Against this normative background, we articulate our proposal in three main parts—principles and values, rights, obligations and recommendations—declined on the individual, social, and global levels. With this work we hope to contribute to spreading awareness on the many social challenges posed by AI and encouraging the establishment of good practices throughout all the relevant social areas.

✉ Andrea Loreggia
andrea.loreggia@unibs.it

Francesco Corea
corea.fr@gmail.com

Fabio Fossa
fabio.fossa@polimi.it

Stefano Quintarelli
aiandsociety@quintarelli.it

Salvatore Sapienza
salvatore.sapienza@unibo.it

[1] Independent, London, UK
[2] Politecnico Di Milano, Milan, Italy
[3] University of Brescia, Brescia, Italy
[4] Copernicani, Milan, Italy
[5] University of Bologna, Bologna, Italy

Published online: 25 May 2022

The present article exhibits three main points of novelty with respect to similar works. First, it rests on an integrated view that combines both human rights and ethical concepts to reap the benefits of the two approaches (Canca 2019). Second, it proposes innovative recommendations, such as those on redress and governance, which add further insight to the debate. Thirdly, it incorporates a specific focus on the Italian Constitution, thus offering an example of how core legislations of Member States might further specify and enrich the EU normative framework on AI.

We acknowledge this last point to be particularly important. The interplay between national and international views is a delicate but necessary step towards the adoption of shared normative frameworks on AI. Due to cultural heritage, national laws, and social guidelines, diversity is to be expected to some extent. The meanings and values associated with social and ethical requirements change from time to time and from region to region (see for instance Awad et al. 2018). For obvious reasons, this situation is particularly critical for Member States of the EU. It becomes thus extremely important to study how frameworks that are already in place at the EU level can be further developed at state level in accordance with the constitution of member states. According to the EU's high-level perspective, this process of specification enables for shared governance of technology across varied domains while also allowing for governance at the state level that is most appropriate for commonality. By anchoring the EU framework to the Italian national level, the article offers an example of how this interplay might be envisioned and pursued.

In the light of the far-reaching goals of this paper, we adopted a cross-disciplinary perspective that has required a certain diversity among authors' research and personal backgrounds, which include computer science, economics, philosophy, politics, entrepreneurship, and law. As known, AI is an intersectoral domain that covers a wide range of issues, ranging from labor to law, ethics, technicalities, economics, and so forth. Our approach mirrors the one pursued by cross-disciplinary working groups, such as the EU High Level Expert Group on AI or the Italian National AI Strategy (Agenzia per l'Italia Digitale 2018), yet bearing in mind the scientific nature of this contribution.

This paper is organized as follows: In Sect. 2 we take into consideration the most fundamental documents of our social organization, with a particular focus on the Italian Constitution. These documents constitute the theoretical foundations of our approach and the normative architecture that technological innovation must be aligned with. Moreover, in Sect. 3 we identify the most important ethical principles rooted in our culture, which provide ethical reasoning with relevant orientation and insight in facing issues concerning the development of AI. In light of these results, in Sect. 4, we specify our general framework in a system of rights that are particularly relevant for innovation in AI. We strongly believe that human life can truly flourish only if all its aspects are duly acknowledged. Finally, in Sect. 5, we report a set of practical obligations and recommendations for the development of applications and systems based on AI techniques are drawn from the outlined premises.

2 Preliminaries

With the aim of encouraging ethical and socially beneficial innovation, researchers, policy-makers and organizations have defined lists of principles and guidelines to steer research efforts and industry towards a common development of the technologies. Particularly, we paid attention to the processing of what is defined by the following works:

- Partnership on AI (Partnership on AI 2016): list of principles focused on the need to develop a culture of cooperation among researchers in AI, to guarantee distribution of the benefits of new technologies as fair as possible and the involvement of public and corporate stakeholders. Promoted by the main Over The Top.
- The Asilomar AI Principles (Future of Life Institute 2017): manifesto of roboethics principles and guidelines for the development of new technologies, defined by academics and professionals in the sector.
- AI in the UK (House of Lords Select Committee 2018): a study carried out by the House of Lords in support of the social sharing of the benefits deriving from the use of a transparent and secure AI.
- Villani (Villani et al. 2016): report by the French parliament defining its strategy on Artificial Intelligence, listing the fundamental principles for its development.
- AI4People (Floridi et al. 2018): the document, prepared by Atomium-EISMD, clarifies the risks and opportunities that AI presents to contemporary society and outlines ethical principles to which to adapt research and use of AI.
- CEPEJ (European Commission for the Efficiency of Justice (CEPEJ) 2018): the document, drawn up by the European Commission for the Efficiency of Justice, aims to assess the ethical impacts and potential of the use of AI in judicial contexts.
- HLEG-AI Ethics Guidelines (European Commission High Level Expert Group on Artificial Intelligence 2018): guidelines defined by the group of experts of the European Commission for the creation of reliable and trustworthy artificial intelligence.
- IEEE Ethically aligned design (IEEE 2017): establishes that the technologies must incorporate, through practices to be implemented already at the design stage, the funda-

mental values to which to associate corresponding policy provisions and legal frameworks.

- Artificial Intelligence Act (AI Act) Proposal :[1] On 21 April 2021, the European Commission proposed the first legal framework on AI to address the risks posed by this method of computation. Under this forthcoming piece of legislation, the EU aims to strike a balance between fostering the EU market, encouraging investments in AI systems, ensuring product quality, and mitigating threats to fundamental rights and freedoms.

We have chosen to anchor the work to some fundamental documents of our social fabric, starting from a global level down to a national one. In the next paragraphs, we will give a brief description of the foundations of this work.

2.1 Sustainable development goals

The Sustainable Development Goals (SDGs)[2] have been formulated by the United Nations in 2015, with the aim of promoting sustainable development through the solution of some of the major economic and social problems of humanity; the development of AI can be beneficial to their achievement (Cowls et al. 2020). In this work, we will discuss the benefits offered by AI to foster an inclusive production process that respects human labor, as set out in goals no. 8 (i.e., promoting sustained, inclusive, and sustainable economic growth, full and productive employment, and decent work for all) and 9 (i.e., building resilient infrastructure, promoting inclusive and sustainable industrialization and fostering innovation).

We shall discuss the importance of distributing the benefits of the technology for reducing inequality within and among countries in accordance with goal no. 10. The promotion of gender equality set in goal no. 5 requires to guarantee the elimination of bias from the design of the algorithms; the protection of equal education is defined in goal no. 4. It demands that digital culture must be spread at all levels of education. On the political level, the promotion of peace, justice, and strong institutions envisaged in goal no. 16 increasingly depends on ethically sound use of the AI in the personalization of mass communication, in the prevention and repression of crime, and in the administration of justice, without falling into forms of manipulation and authoritarian state control. With regards to the promotion of peace, in particular, AI must in no way replace human judgment in weapon systems.

2.2 Universal declaration of human rights

The Universal Declaration of Human Rights[3] constitutes one of the essential documents for every ethical–juridical reflection from which to start also a dialogue on the ethical framework of the AI. This dialogue is based on the recognition of human dignity as the foundation of respect and promotion of rights, regardless of gender, ethnicity or religious affiliation, political opinions or other factors that may give rise to discrimination. The universal character of the human rights recognized in the Declaration makes it suitable to favor a global and inclusive discussion on the themes and challenges launched by the AI regarding the contemporary pluralist and multicultural society. Of particular relevance, for the purposes of this document, are dignity (art. 1), protection of privacy (art. 12), freedom of information (art. 19), and attention to issues of equality and non-discrimination, fundamental in the reflection on algorithmic bias. The member states of the United Nations are subject to the obligations arising from the Declaration, which must transpose its principles into national laws concerning the subjects regulated by them, not directly bound by the Declaration.

2.3 Charter of fundamental rights of the European Union

The Charter of Fundamental Rights of the European Union establishes fundamental values and objectives of the European Union (Chalmers et al. 2019) and represents a relevant reference framework for the development and use of AI systems. The values outlined are based on inclusion, tolerance, justice, solidarity, and nondiscrimination and are expressed with respect for the dignity and human rights, individual freedom, democratic ideals, equality of citizens before the law and of the rule of law. In light of these values, the EU is committed to promoting peace and well-being, freedom and security for its citizens; to guarantee justice and freedom of movement; to promote sustainable development based on balanced economic growth and price stability, on a highly competitive market economy, with full employment and social progress, and environmental protection; to fight against social exclusion and discrimination; to promote scientific and technological progress; to strengthen economic,

[1] Proposal of the European Commission for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, adopted on 21 April 2021 (COM (2021) 206 final).

[2] The complete and official list of SDGs is available online—https://sdgs.un.org/goals—last visited on March, 18th 2022.

[3] The Universal Declararion of Human Rights is available on line—https://www.un.org/en/about-us/universal-declaration-of-human-rights—last visited on March, 18th 2022.

social and territorial cohesion and solidarity between the Member States while respecting the cultural and linguistic diversity that characterizes their nature.

2.4 Constitution of the Italian Republic

Amongst its fundamental principles, the inviolability of human rights (Rodotà 2015) and the recognition of formal and substantial equality must be guaranteed in the research and use of AI systems. Furthermore, the centrality of work (art. 1) requires a reflection on the economic and social consequences of the development of such systems. In the application of weapon systems, it emphasizes the rejection of war as a means of resolving international disputes (art. 11). Respect for citizens' rights to personal freedom (art. 13), privacy (art. 14 and art. 15) and free expression of thought (art. 21) must guide the use by public and private entities of AI systems capable of threatening these fundamental freedoms. For these subjects, respect for human dignity in the economic initiative (art. 41) should be recommended. In the activity of safeguarding security and public order, as well as the activities of prevention and repression of crimes, it is necessary that the attribution of civil and criminal liability for the improper use of AI systems is personal (art. 27), as well as strengthening respect for the principle of legality and due process (art. 111).

3 Principles and ethical values

Based also on the analysis carried out in the preceding paragraphs, we have identified a set of ethical–social principles and values, organized in three macro levels with an increasing stratification, from the individual level to the global one. The three macro-categories, and the related principles, are to be understood according to an integrated perspective, not of relevance. The choice of such an approach does not reflect formal needs, such as the compilation of an aggregate of statements, but underlies a will of a substantive kind, i.e., to propose an organic vision.

3.1 Individual level

The individual level aims to identify the values that pertain to the person and are based on the monolithic foundation of human dignity. From the latter derive civil rights and the principle of non-discrimination, all of which are expressed both in the material dimension and in the immaterial dimension of human activities.

3.1.1 Human dignity

Although it is extremely difficult to agree on its definition, the principle of human dignity is widespread and commonly recognized. As such, it plays a fundamental role in the Universal Declaration of Human Rights, in the Charter of Fundamental Rights of the European Union and in the Constitution of the Italian Republic. In its most fundamental sense, dignity corresponds to the intrinsic value that pertains to every individual as a human being—a value that, to refer to a well-known Kantian formula, requires never treating another human being only as a means for your own purposes, but also and always as an end in itself, that is as a subject able to determine himself/herself independently. The principle of human dignity constitutes a limitation of the power of self-determination and action of the individual, so that the intrinsic value of his or her counterpart acts as the boundary of his or her own freedom. Given its social pervasiveness and the profound impact we believe it will have on every aspect of life, AI could have significant effects on respect for human dignity (European Commission High Level Expert Group on Artificial Intelligence 2018; Floridi et al. 2018). Applications in the industrial, health, educational, welfare and social fields will be able to offer new powerful means for the production, maintenance or strengthening of the conditions associated with a dignified life. However, the same technologies that can be indicated as means for respecting and affirming human dignity could also threaten both moral and physical integrity, intersecting the bioethical themes of transhumanism. Contrary to human dignity, they seem to be technologies that manipulate the user—even for good—or to which decisions of great social or existential importance are delegated without being possible to understand the dynamics. Again, human dignity is significantly at risk from technologies that do not capture the intrinsic value of each individual by dissolving its particularity in the generality of statistical models. In conclusion, the principle of dignity is widely recognized as a fundamental requirement for the development and ethical use of AI.

3.1.2 Freedom and civil liberties

The dignity of the person is the principle on which the ethical model of civil rights is based. The most reliable way to ensure happiness and justice is the affirmation of the human being value that differentiates it from other natural beings and also gives it its dignity. It can be said that the Universal Declaration of Human Rights is considered the origin and the fundamental nucleus of an ethical construction that works as the basis of the solution to the conflicts of human coexistence. The Universal Declaration of Human Rights is material ethics that establishes values, contains rules that must be respected, rights that must be guaranteed and freedoms that must be protected. Historically, the first developed rights have been the so-called first generation rights, freedom rights, which limit the power of the state, such as freedom of thought, conscience and opinion, in response to absolute monarchies and dictatorial regimes. Subsequently, the rights of second-generation concerned the rights of equality and political rights, which ensure a level playing field in participation in political power. The third generation, whose fundamental value is solidarity, includes social rights, the right to social security and the right to work. As we see, human rights must be viewed in a dynamic perspective: they have evolved over the course of historical experience, and it is reasonable to believe that they will continue to do so. Civil rights are rooted in the Universal Declaration of Human Rights which recognizes the right of all people to freedom (of movement, thought, opinion, association, etc.), to justice, an adequate standard of living, to health and well-being, in particular medical care and social services. All of these are areas in which technologies take on a predominant role. Thanks to AI, the immaterial dimension has in fact become (or is becoming) the main user interface for the social and economic relations of people, the first place in which these rights must be insured (Quintarelli 2019). This assurance must be substantial, not just formal, thus balancing the existing gaps in the material dimension between different individuals and including a particular caution for the weakest people who statistically would be relegated to outliers in the statistical models.

3.1.3 Non-discrimination

The Charter of Fundamental Rights of the European Union affirms the principle of equality: to recognise all citizens the same rights before the law (Chalmers et al. 2019). The principle of equality between men and women is the basis of all continental policies and is the element on which European integration is based. It applies in all sectors. The Italian Constitution states:”All citizens have equal social dignity and are equal before the law, without distinction of sex, race, language, religion, political opinion, personal and social conditions. It is the duty of the Republic to remove those obstacles of an economic and social nature which constrain the freedom and equality of citizens, prevent the full development of the human person and the effective participation of all workers in the political, economic and social organization of the country”. In addition,”The Republic recognises all citizens the right to work and promotes the conditions that make this right effective”. From these anticipations follows the fight against discrimination, the protection of minority rights and the most fragile sectors of the population in relation to their objective situation. The data collected and used in machine learning systems describe the social fabrics incorporating the related prejudices. In the absence of specific precautions and provisions, statistical models materialize and possibly amplify these biases (Jobin et al. 2019).

3.2 Social level

AI systems, being non-deterministic, tend to generate some incorrect and ethically problematic predictions. In assessing their consequences, conflicts might occur between the individual level and the social level. This section considers relevant values that should be respected and promoted with a view to striking a balance between social and individual well-being.

3.2.1 Inclusiveness

From the point of view of fairness, the adoption of AI applications must guarantee fair access to opportunities, services and work to everyone, with a special eye on weak or disadvantaged categories. Furthermore, concentrations of resources and power should be avoided. The effects of AI-based predictions and decisions on inequalities concern not only economic aspects but also wider issues pertaining to the socio-cultural context (O’Neill 2016). In this sense, financial difficulties might overlap with other concerns, such as, e.g., lack of access to digital education and discrimination on the basis of ethnicity or gender. A prejudicial design and use of new technologies risk further undermining the condition of social groups that are already in precarious situations, thus triggering a vicious circle of marginalization and a further increase in inequalities (Eubanks 2018). A fair adoption of smart technologies requires an inclusive distribution of opportunities (be they financial, educational, legal, health-related, welfare, and so on). AI tools, services, and the associated benefits must be accessible to as many citizens as possible, regardless of their social status, class of income, geographic location and other similar factors.

3.2.2 Inequality reduction

AI should be developed to prevent and actively reduce inequalities, ensuring maximum sharing of the socio-economic benefits of the new technologies. The productivity gains guaranteed by its implementation should not become the monopoly of a limited circle of subjects, but should instead be distributed fairly across different categories and social classes. AI can become an active force for the reduction of inequalities (e.g., (Capucha et al. 2020)), incorporating a concept of distributive justice that looks at marginal categories as subjects of priority intervention. Tools based on AI, for example, can be useful, within the educational system, to bridge learning gaps (Drigas and Ioannidou 2013; Kharbat et al. 2021), while in the health care sector, they can be used to stimulate social empowerment and services for individuals with disabilities (Neto et al. 2019). From the perspective of the Italian Constitution, the impact on work, employment and wages primarily requires due attention. In this respect, measures should be taken to ensure that negative phenomena (Pham et al. 2018), such as mass redundancies, generalized unemployment, de-skilling and the depreciation of human labor are dampened appropriately. Moreover, new labor policies to address the challenges brought about by automation and new social devices capable of both mitigating its negative externalities and favoring generalized conditions of dignified existence should be elaborated.

3.2.3 Social cohesion

The development of AI must foster social cohesion and ensure the robustness of the democratic process. Previous studies (Sirbu et al. 2019a) have shown how AI systems can help members of a community to reach consensus through a reduced number of interactions but can also lead to strengthening and root divisions over time (Sirbu et al. 2019b). This effect, known and exploited by mass media, assumes an extremely significant relevance in the age of personalized media. In this context, socially desirable goals and legitimate business interests can diverge and conflict with each other. On the one hand, the collective interest in strengthening social cohesion suggests the adoption of technologies capable of favoring the composition of different opinions and promoting tolerant comparisons (Loreggia et al. 2020). On the contrary, to maximize user engagement, the number of interactions, and, therefore, screen time and associated revenues, companies are led to adopt AI technologies aimed at amplifying divisions and exacerbating spirits, features exploited for the spread of so-called fake news (Lazer et al. 2018) and deep fakes (Vaccari and Chadwick 2020). Another way in which this divergence of objectives is expressed is the use of algorithms that are able to exploit user confirmation bias by hyper-customizing messages. The targeted re-proposal of similar content (so-called echo chamber), justified by the need to improve the user experience, risks compromising information pluralism (Cinelli et al. 2021). These phenomena, if not addressed, negatively affect democratic processes and undermine social cohesion with profound and long-term socio-economic effects.

3.3 Global level

In continuity with the principles discussed at the individual and social levels, damage prevention, the search for peace and justice, and sustainability are finally configured as global cornerstones of the ethical development of AI.

3.3.1 Damage prevention

Computer systems allow users to tackle problems with substantially unlimited scalability, well above that possible for humans. Moreover, AI systems allow to face problems of a different nature with respect to the traditional application domains of deterministic algorithms. Consider, for instance, perception and classification tasks previously reserved to human activity, which can be the object of a substantially infinite scalability at a global level. On the one hand, extended scalability increases human possibilities. On the other hand, it expands the purview of possible risks. Indeed, the use of technologies is as global as the propagation of errors, biases, and relative harms. Damage prevention takes the form of a risk assessment aimed at adopting or applying measures that prevent harm from occurring or mitigate its exposure or effects. Prevention becomes a dynamic process that periodically evaluates systems through risk assessment procedures and protocols for risk management. The establishment of good practices—such as those introduced in the field of information security (ISO/IEC 27001)—would allow to identify and describe critical situations. Risk management practices adopt and promote "risk scenarios" as a useful methodology for risk analysis. It is of increasing interest to collect and prototype scenarios that can be used to assess autonomous systems to be able to identify different risk classes (McGregor 2021). In such a way, risky situations (such as the transmission of bias to the data used for training, data poisoning, and adversarial attack) can be surveyed and used for a periodic evaluation of exposure of existing systems (Vedder 2019).

3.3.2 Peace and justice

Article 3 of the Treaty on European Union establishes the promotion of peace and justice as shared objectives by the Member States, in line with the European Convention on Human Rights and the Charter of the United Nations. Article 11 of the Italian Constitution also commits to ensuring peace and justice among Nations. As seen in the section dedicated to Social Cohesion, the ICT revolution may result in accentuating inequalities within the developed countries. In addition, it might dig an even deeper gap between these and the developing countries (Alonso et al. 2020). In some cases, this might require additional considerations. For instance, in projects that involve less advanced or developing countries, it is essential to consider if and how AI systems can be integrated with the solutions already adopted in these contexts and what resources are necessary for their effective implementation. In particular, the scarce quantity and quality of digital information gathered at the intervention areas can hinder the adoption of AI systems. In addition, a reflection is needed on the possibility of making semantically interoperable data available to the least developed or developing countries collected or developed by technologically more advanced countries. An ethical development of AI must, therefore, guarantee the protection of social values, such as international harmony and human brotherhood. Furthermore, technological innovation must be conducted in compliance with the principles of justice and help prevent escalation of international conflicts and tensions (Haner and Garcia 2019; Taddeo and Floridi 2018b). In the prevention and repression of crimes, the principles of legality and fair trial—which is guaranteed by international law, enshrined in the Constitution, and recognized by multiple jurisdictions—must be placed as indispensable prerequisites for the deployment of AI systems in these sectors.

3.3.3 Sustainability

Sustainability—in its environmental, economic, and social aspects—poses significant challenges to the future of humanity and calls for swift global countermeasures. Its importance is exemplified by the already introduced 17 sustainable development goals (SDGs), which serve as a framework to organize international action. The digital revolution, spurred by continuous progress in the field of information technology, machine learning, and robotics, is likely the biggest single technological factor that might help achieve the SDGs in the coming years (Goralski and Tan 2020; Vinuesa et al. 2020). Consider, for example, its potential impacts on the economic aspects linked to sustainability. AI will increasingly affect almost all sectors of the economy. Agriculture (precision agriculture), mining (autonomous vehicles), production (robotics), marketing (profiling), finance (models behavioral), media (individual targeting), health (diagnostics), etc. are all experiencing major transformations thanks to digital technologies. In general, these technological contributions can increase productivity, reduce production costs, reduce environmental impacts by dematerializing production, and improve market functioning. In a word, the digital revolution has the potential to offer significant support to the many sustainability challenges to be faced in the next future. However, there are also obvious risks that must be identified and addressed. Among these, the emergence of monopolies requires due attention. In fact, some companies can exploit their advantages in accumulating large data to gain a dominant monopoly position in their respective markets, allowing them to enjoy a sheltered position from competition and altering the regular functioning of the market. Another most feared concern is the loss of jobs and the displacement of income distribution from labor to capital (Korinek et al. 2019). Apparently, an important consequence of automation processes is the reduction of the demand for less qualified workers (Arntz et al. 2016). With advances in AI and robotics, many more workers can now see their jobs and incomes threatened (Frey and Osborne 2017). While new jobs could replace old ones, they could bear lower incomes and impoverished working conditions. Only by raising awareness and taking action to mitigate such risks will it be possible to fully ripe the benefits that the digital revolution might yield in terms of economic sustainability. In addition, from the social standpoint it will be crucial to strike sustainable balances between benefits and risks. Social advantages brought about by digital technologies are many and range from expanding access to goods and services to enhancing pharmacological research and therapies, reducing waste, allowing for a better management of public resources, and simplifying access to public services (Taddeo and Floridi 2018a). There are, however, many other social threats perceived as brought about by the digital revolution. Digital identities can be stolen. Governments and private companies can invade privacy and monitor individuals against their will or without their knowledge. In addition, social media can be manipulated and cyber-attacks can paralyze companies or institutions by interrupting information flows or hitting devices connected to the Internet. Finally, the problem of the environmental impacts of AI technologies requires special attention as well (Nishant et al. 2020). Even though in many cases AI can help optimize processes and, thus, reduce waste and emissions, it is important to raise awareness and take into due consideration the high energy consumption and the extraction of materials for chips or batteries that are necessary to power, train, test and deploy AI systems (Lucivero 2020; Vinuesa et al. 2020). Moreover, research must be carried out to predict possible rebound effects and take action to minimize their effects.

4 Rights

This section focuses on rights deriving from previously analyzed ethical principles and values. These rights, we suggest, can be drawn from the background against which recommendations about the ethical development of AI technologies are benchmarked. Discussing these rights in the light of AI systems is necessary for a twofold goal. On the one hand, a careful re-consideration of their nature, scope, and applicability is crucial to measure the impact of AI systems on fundamental liberties; on the other hand, such discussion contributes to the ongoing regulatory process of AI systems and, ultimately, to ensure that emerging technologies are compliant with the constitutional prerogatives guaranteed by fundamental treaties and charters.

4.1 Information

All autonomous systems require individuals to have an active exchange of information to provide a service or suggest a solution to the problem faced. It is necessary to ensure that user rights are upheld during interaction. It is a user right to know and be informed about the entire process (Pagallo 2020; Palmirani 2020): from the collection of data and information to the processing procedure, from the risks to the very nature of the interaction with the system (whether this involves an autonomous system able to process information or not). Informed consent should be presented in a clear and succinct manner allowing a conscious choice and avoiding impulsive or nudged acceptances, including by means of design choices in the presentation of the information set (Rossi et al. 2019). In the case of decisions that can have a significant impact on users' lives or on society as a whole, it is necessary to protect the right of free choice with regards to the desired level of autonomy/intelligence of the system during the interaction. For example, the consequences of such choice should be specified and individuals subject to the decision should be granted the possibility of requesting the intervention of a human operator (Art. 22 of the GDPR).

4.2 Education

Awareness of opportunities and risks related to technology goes hand in hand with education and technological training. It is desirable to educate, instruct and train social groups and individuals for the proper use and mature coexistence with technology. We propose to conceptually differentiate education, instruction, and training to analyze important aspects of the human–machine relationship separately. In this context, educating means knowing how to frame relationships between people and technology—in particular, how the individual should relate and interact consciously with the provided tools. Instructing means knowing how to transmit the knowledge that allows the person to know (even in a summary or general way) how the technology works and consequently to assess its risks and potential. Training refers to a learning process through which users (aware of their knowledge and gaps) improve and increase their education. In this perspective, the abuse of technology becomes a lack of education, while the proliferation of catastrophic alarmism is a lack of instruction on technology (the "doom-monger position" in Agenzia per l'Italia Digitale (2018)). For this reason, raising the level of information literacy would result in greater adequacy and awareness of individuals, making them more suitable for facing the rapid evolution of the world. In a world that is increasingly populated by autonomous systems, being able to interact with them consciously and appropriately should be considered as a citizen right and public action should be taken to uphold its demands.

4.3 Self-determination of identity

We propose to use as a definition of a person's identity the set of material and immaterial attributes that define him or her, describing his or her uniqueness and diversity (Floridi 2014). Self-determination of identity is an unavoidable and inalienable right rooted in the dignity of each individual (Rodotà 2015). The social nature of human life makes personal identity a social characteristic, that is shaped through a path, where free interactions with the environment and other individuals allow us to build a narrative that changes over time and according to the environment (Rachels 1975). Data collection and interaction with technology in everyday practices make artificial intelligence a powerful tool capable of performing tasks and satisfying needs by improving the quality of life, but at the same time it can become a means to manipulate the decisions of individuals, thus undermining their self-determination (Taddeo and Floridi 2018a). The same area also includes more pragmatic aspects of identity protection, such as the right to portability, rectification, erasures and other related rights. To make these forms of protection workable, the individual must always be in a position to know the nature of its interlocutor (artificial or not), its aims and potential, to allow him or her to choose how to interact with the agent and which faculties to grant it on his or her data.

4.4 Privacy

Departing from the seminal work of Samuel Warren e Louis D. Brandeis (Warren and Brandeis 1890), the protection of privacy has been linked to the recognition of a private sphere within which the individual shall be immune from interference by third parties, be they public or not. This reading has been also endorsed by the Italian Constitutional Court in cases referred to art. 15 (Rodotà 2015) and in European courts (Lynskey 2014). The increase in computing capacity, storage and connection determined by the evolution of electronics has led to the accumulation of data over time. This information originates from devices and sensors deployed both in the material environment, where individuals live and in its immaterial counterparts in which they project their identities; the increase of the processing capacity, thanks to the statistical models that are deployed, determines a shift from applications that rely on deterministic algorithms to probabilistic approaches. Therefore, three dimensions whose relevance grows exponentially with respect to traditional computer applications call for specific attention: spatial interpenetration, temporal accumulation and statistical modeling. The first dimension requires a conceptual re-engineering of the spaces within which individuals move and live (Floridi 2015). With the gradual erosion of the barrier that separates immaterial and material dimensions, it is necessary to evaluate the impact of AI systems in the private sphere of the individual considering the interconnection between the physical environment in which his or her movements take place and the digital perception that AI systems have of them. A second perspective revolves around time as a distinctive element of analysis. The use of data related to the past for the development of probabilistic predictive models calls for a reflection on the possible effect of social crystallization that could happen if the most invasive intrusions in the privacy sphere were taken without assessing the need of correcting, integrating or eliminating bias or data that are no longer relevant. In the cases of permanent and indiscriminate accumulation of data ("always-on"), it is necessary to consider which possibilities are offered to individuals for reasons of privacy to escape, even temporarily, from the collection of personal information. A third perspective is linked to the cognitive profiles of AI and to the extraction of new knowledge from data. In the assessment phase of the pervasiveness of AI systems in the private sphere of the individual, emphasis must be placed not only on the management of the observed personal data, but also on the impact of the inferred information that these systems are able to generate and on the super-individual perspective through which they allow us to observe reality.

4.5 Protection of rights

Effective protection of one's rights is itself a fundamental right of people in line with the objective of the SDGs and the need to ensure effectiveness of the essential conditions of democracy (Rodotà 2015). The aforementioned recent developments of electronics make them pervasive and capable of capturing data from every action and interaction. It is possible to observe a shift to pervasive devices in which functions and personal data are distributed and access to these functions and data is ensured through simplified human–machine interfaces based on identity recognition. Thanks to such availability of data and computational capacity, AI techniques allow the creation of software that is no longer algorithmically deterministic but based on the application of statistical models extracted from data. The nature of computer applications changes accordingly, thus allowing the typical scalability of computer applications to be applied to different areas that cannot be addressed with previous deterministic algorithms, including, for example, cases traditionally bound to human perception and their classification. In these uses of AI, the scale factor changes the very nature of the application. Let us consider the passage of the examination of police photographs (mugshots) carried out by human beings and, therefore, limited to a few tens of thousands of individuals, to its execution by means of AI instruments, indefinitely scalable, and, therefore, potentially concerning millions of people. In defining statutory reserves, the Italian Constituent Assembly could only envisage "human" interventions which implicitly incorporate frictions and limits to scalability. With the scalability offered by AI systems, the nature of control and the degree of interference can change from an exception to a social rule that cannot be considered as a simple technological upgrade but that raises new questions about the protection of individual rights. Despite the changing of informational context and the ongoing dematerialization of users and media, the magnitude of AI scalability should not deprive us of the critical capacities that only reading the Constitutional charters can guarantee.

4.6 Rights of fragile individuals

Some categories of individuals (such as minors, elderly and non-self-sufficient people) can be found in the inability to make autonomous decisions, incapacity linked to age or psycho-physical conditions. They fall within the protection granted by the fundamental rights of 'social solidarity' in the Italian Constitution (art. 2). The UN Convention on the rights of children and adolescents establishes in particular how the minor is entitled to essential rights, as well as a subject of special care and assistance.[4] Art. 29 sets forth the importance of the development of personality, identity and attitudes with special references to the environment and to interpersonal relationships described through different levels of cultural interchange. Technology and in particular AI systems are instead designed and built to capture the attention (especially of the young individuals (Paul et al. 2012)) and to keep the latter as long as possible to collect or produce more informational material possible. Technology should instead consider these categories as targets of a stronger protection, thus promoting values and methodologies that place them within an environment of assistance and growth by favoring their cognitive development and allowing its free determination, without binding choices, preferences and attitudes, in particular when such manipulation is instrumental to commercial exploitation or for manipulative purposes (SDG 3).

[4] United Nation Convention on the Rights of the Child Adopted and opened for signature, ratification and accession by General Assembly resolution 44/25 of 20 November 1989.

5 Obligations and recommendations

In light of the principles and rights identified above, this section focuses on identifying obligations and proposing recommendations to be followed in the ethical development and the regulated use of AI systems.

5.1 Trust

Trust plays a crucial role in every innovation process: only by protecting and promoting the social trust capital relative to AI it will be possible to fully grasp its potential (European Commission High Level Expert Group on Artificial Intelligence 2018; Floridi 2019). Therefore, it is necessary to identify the factors that could undermine trust in this technology and implement effective measures that minimize or erase negative externalities. Trust is a primary social bond and supports the organization of work, sharing tasks and the delegation of jobs, making opportunities and perspectives otherwise unattainably available at a community level. Given the importance of trust in every human interaction, it is necessary to take precautions so that also the technologies based on AI—as mediators of social relations—are reliable, worthy of the trust of the different actors involved in their use (SDG 9). The developed technologies must reflect the values of the users and of the society in which they are deployed, in accordance with socially shared objectives. Trusted frameworks, *corpora* of methodologies, rules, certifications, controls, sanctions, benchmarks aimed at achieving socially desirable and politically determined targets must, therefore, be carefully identified and applied.

5.2 Accessibility

Since the social impact of the decisions taken by means of algorithms is considerably relevant, it is of great concern that the ways in which machine learning techniques derive and infer information are not easily accessible, that is, they are not transparent and explicable. This is the reason why it is recommended by many voices that autonomous decision-making systems are as intelligible as possible and that their determinations can be explained. The opacity of the ways these systems process the data on which the decisions produced are based—which makes them black boxes (Pasquale 2015)—is problematic both from a technical and a social point of view: the lack of accessibility can lead to the suspicion that some correlations on which the decision is based can incorporate ethically reprehensible prejudices and cause discrimination, unfair treatment and injustice (Mittelstadt et al. 2016). On a technical level, transparency is necessary to make the system's decision-making process and its internal logic knowable and to be able to validate it from a technological point of view. On a social level, explicability refers to the translation of the algorithm function into terms understandable to users or those subjects to the decisions, to provide the reasons for the output. In fact, one can have explainability without transparency, e.g., when a decision is taken on the basis of a known criterion but without access to the algorithm that processes it. We can have transparency without explainability, where full access to data and algorithms is provided but the determination of the output is not sufficiently motivated by the system. The value of transparency requires the development of new explanation techniques capable of opening black boxes and accounting for their internal processes. The guarantee of accessibility is inseparable from the elaboration of policies which, by including transparency and explicability, strengthens trust and protects the right of users—citizens, authorities and the scientific community—to be informed simply and clearly about the use of artificial intelligence and the limits related to such use. Accessibility becomes particularly relevant when framed within the discussion related to the use of machine learning systems in the judiciary, in particular when judges are obligated by Constitutions to provide reasons for their decisions to allow challenges and appeals, e.g., in the case of Article 111 of the Italian Constitution.

5.3 Safety

The protection of personal safety derives directly from the fundamental values of personal dignity and as such it requires it to be respected at every stage of the process of designing and using AI technologies. As a result, we focus on safety as a system integrity. The pervasiveness of the technology makes safety an essential obligation that the supplier, of services or technology, must ensure at different levels of applicability: from the protection of the safety of individuals to the storage of personal data, from the protection and management of physical assets to system structural integrity. Regarding personal data, we recall the provisions of the EU Regulation 2016/679.[5] Among the other rights guaranteed by the legislative text, it is necessary to provide suitable measures to allow the limitation of the processing of personal data. This faculty takes on a particular significance of protecting the safety of data against their use by computer systems (Rodotà 2015) and, even more so, by AI systems. A safe system should then exclude and avoid possible negative externalities, as well as be structured in such a way as not to

[5] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46.

incentivise the system itself or third parties to achieve the goals set by unsuitable tools or actions. At the same time, the system should be allowed to learn new strategies to complete an action without such having unexpected repercussions, and to ensure a degree of flexibility such that it can adapt to different situations without having to be checked at every step, especially if the mechanisms monitoring systems are complicated or expensive. Identifying accurately the safety level of a system is as important as making sure that such level is communicated to the users in an intuitive and clear way. From this point of view, a functional language should be developed—for example, through the use of certifications or labeling mechanisms—which simplifies the understanding of the level of safety and reliability of an AI technology.

5.4 Usability

As highlighted above, a significant number of AI systems are based on complex computational models, occasionally not very intelligible and opaque (black box (Pasquale 2015)), which can generate a perception of lack of control in the execution phase. If, on the one hand, the relevance of the intermediation between system and human beings is justified by the need to ensure transparency, control and trust (Amershi et al. 2019), on the other hand, it acquires even greater significance when special categories of users make use of the system. The benefits offered by AI systems shall be accessible to people who are not self-sufficient, to ensure the maximum expression of their potential and a meaningful life (Article 3 of the Italian Constitution, Article 26 of the EU Charter of Fundamental Rights).

Special attention must also be paid to the conscious use of AI systems by minors. It is thus necessary to discuss usability of AI systems according to a twofold perspective: on the one hand, on how to adapt them to the different stages of the child's development (for example, development of linguistic skills or mathematical thinking) and to their subjective conditions (cultural and language background or learning disabilities); on the other hand, on what limits shall be placed on the development of manipulative interfaces or capable of generating confusion about the artificial nature of the system (SDG 4). It is, therefore, necessary, for a correct interaction between humans and AI systems, that interfaces place human users at the center of its development project, in particular when they belong to particularly sensitive targets due to objective conditions.

5.5 Control

An active participation of a human being in the decisions taken by AI systems is necessary so that the operator does not have the role of passive executor exempt from moral and legal responsibilities connected to the use of such systems. This supervision prevents the reduction of the recipients of decisions to mere variables of a probabilistic calculation, an unacceptable condition in the case of critical situations in which there is a risk for ethical values of constitutional relevance or where a moral evaluation of the consequences of the decision is necessary. The discussion on the control of AI systems can be divided into two directions: a descriptive one, in which the degree of autonomy of an AI system is defined (for example, through an ordinal quantity in which each value corresponds to a certain degree of self-determination of the machine and, conversely, the degree of human control (SAE International 2018); a regulative one, in which each level of control is translated into a given legal regime, thus allocating different and appropriate liabilities depending on the context of use. Therefore, we recommend to comply with the human-in-the-loop principle and stress the importance of fostering human active participation in algorithmic decision processes.

5.6 Accountability

The Accountability dilemma is certainly one of the most problematic aspects in the development of new AI systems (Villani et al. 2016). However, it is not clear whether responsibility for certain decisions made by an intelligent system should be attributed to its developer, the software vendor, the user or third parties. Furthermore, it is interesting to note how different people have developed a natural and irrational "aversion to algorithms", which introduces an additional element of responsibility: if a doctor decides not to follow the recommendation of an AI system that is not considered reliable, but is wrong, can the doctor be held responsible for the outcome of this decision? How far can we ignore an AI system? It is clear that such problems are not easy to solve, and that it is probably difficult, if not counter-productive, to create and use a single reference system to manage the Accountability of different actors in different circumstances. In fact, multiple frameworks could be required that consider not only responsibility as such, but also transparency, fairness, accuracy and the degree of control of an algorithm (European Commission High Level Expert Group on Artificial Intelligence 2018). For applications that can have a significant impact on societies, people and things, the responsibilities connected to the use of these systems must be assigned *ex ante* instead of waiting for the *ex post* evaluation of a third subject called to allocate objective responsibilities. The contractual terms that detail rights, faculties, immunities and privileges must be clear and accessible, especially in applications that are intended for or affect a large number of people. Finally, it is necessary to develop accountability mechanisms that prevent strategies of non-responsibility or responsibility being assigned to non-human subjects.

5.7 Redress

As we have the principle of "privacy by design" for personal data management systems, it is appropriate to consider the introduction of repair mechanisms (redress), for AI-based systems that make decisions that can affect people's lives, on the basis of a "redress by design" principle. The basic consideration is that a non-defective, fully functional AI system will make incorrect predictions. This happens both because of the bias that might be present in the training data and because a system of this type is inherently non-deterministic, such as, for example, a speed camera can be: if you exceed the speed limit while driving your car, the speed camera detects it and you get a fine. Despite being possible to appeal this decision, the driver is found guilty until proven innocent, because a deterministic, non-defective (assuming its proper configuration, certification and control) system establishes his or her guilt. However, with a non-defective, fully functional artificial intelligence system, being a statistical engine that necessarily produces probabilistic results, this decision could be right 98% of the time and wrong 2% of the time (it would be inappropriate to classify these incorrect predictions as mistakes), which means that in this 2%, the person is found guilty even when he or she is not (or cannot obtain a service, even if he or she has the full right to obtain it). For the person, the wrong decision can generate relapses, overcoming the scope of the decision itself, for example generating social blame, negative online feedback and other consequences that can spread across the Net and become impossible to remove. This 2% tolerated error is not to be understood negatively, since it guarantees a structural flexibility to the algorithm and the ability to adapt and include new emerging elements. In these wrong cases (they may be either false positives or false negatives), the appeal procedure may not exist or, if it exists, may be ineffective, its cost may be excessive, it may be not accessible to all, it may take too long, or it may not correct the aforementioned relapses. More effective protection of rights should include the principle of redress by design, or the provision, from the design phase, of mechanisms to ensure redundancy, alternative systems, alternative procedures, ombudsman, etc. to be able to effectively identify, verify, correct the wrong decisions taken by a non-defective, fully functional system and eventually, refine the predictive capabilities of the system. This principle can be read in conjunction with access to justice granted by Article 111 of the Italian Constitution and the 'right to a fair trial' when automated decisions take place in the context of judicial scrutiny.

5.8 Ownership of data

The data pertains to the individual who generates them. There are different subjects that use data to extract information. The information improves the level of the services obtained in terms of accuracy and reliability. For instance, in modern navigation systems, the aggregated information makes it possible to know the status of traffic in a given stretch of road by informing users and enabling them to change the route, thus improving the level of service provided. This information is fundamental for the definition of mechanisms able to provide useful services to people, to improve their quality of life. The protection of raw data, of its ownership, of maintenance and possession is only a first step not yet sufficient to guarantee the protection of the individual and the community. With the increase of data collected and processed, the need to define technical, contractual and regulatory mechanisms governing the extraction and management of information increases, identifying levels of accessibility, methods of use and disclosure (House of Lords Select Committee 2018). Even given the non-rival and only partially excludable nature of the data, users must be technically provided with full transparency and control of the data collected and processed by AI systems, guarantees that must be ensured at the contractual level. The scenario can radically change with the introduction of cryptographic systems or distributed computing systems (cryptographic network overlays, homomorphic cryptography or distributed AI systems). Their developments must, therefore, be followed carefully, to allow users full ownership of the availability of the data kept therein.

5.9 Governance

The issue concerning the governance of AI does not technically reflect a single problem, but a multitude of different aspects. It is in fact an all-encompassing problem that touches on issues linked to justice, responsibility, from national strategies and policies concerning AI up to "smart" surveillance.

- The consideration that AI determines a scalability that transcends the natural limits to human perception and classification activities, putting pressure on consolidated social processes, suggests the opportunity to consider arranging the introduction of frictions to limit or slow down this scalability in the cases in which this can cause negative externalities (among the others, the aforementioned example relating to mug shots or the spread of fake news that undermine social cohesion and democratic processes).

- From the intrinsically statistical nature of AI, it follows that individual instances of a non-defective system can cause adverse events due to incorrect predictions, while at the same time causing much greater overall benefits than the previous situation. Instances of a driving assistance system may fail, thus causing some accidents, even fatal ones, but overall its use greatly reduces the number of accidents and victims. In some ways it is a situation similar to that of pharmaceutical products.
- AI governance should ensure an adequate identification, measurement and classification of erroneous forecasts caused by non-defective systems, to ensure that they fall within the socially desirable target values established through democratic processes (SDG 16). In certain cases, it may be necessary to conduct system validation and measurement tests of their effects before they are marketed. In some others, validation and conformity assessment procedures should be defined, at least for systems that may have significant impact risks on properties, people and societies, allowing to correctly exclude or attribute wilful misconducts and negligence.

To address these situations, consideration should be given to the creation of an authority or agency responsible for monitoring the dissemination of AI and detecting emerging challenges, providing information to policy makers and applying fines. It should ensure the fulfillment of objectives at system level set by policy makers in relation to AI application classes, based on the effects on individuals and social organizations. To this end, it could issue guidelines and recommendations for AI system developers and assist companies in applying an approach based on risk and impact assessments. Furthermore, this body could ensure coordination with the European Union and other international standardization bodies. Finally, it could benefit from the collaboration of intermediate bodies, such as trade unions and consumer associations.

5.9.1 Training

AI also holds considerable relevance in the education sector (SDG 4). Therefore, it is necessary to identify to what extent it generates opportunities and risks in certain sectors. The social body must be able to access training paths to qualify and re-qualify itself, so that the impact of AI on the world of labor can meet professionals prepared to seize the opportunities and up to the challenges that such technology poses at the ethical and social level, thus avoiding the formation of marginalized groups unable to find their own role in the new working environment[6] [see for instance an overview of relevant scientific venues, journals, projects and resources in the context of AI K-12 education (Kandlhofer and Steinbauer 2021)]. Users' training should allow a conscious use of technology that does not demonize its nature but unlocks its potential by making the individual aware of risks and dangers. The corporate world should promote educational paths aimed at facilitating the integration of ethical considerations related to developing AI technologies, both in design—by supporting interdisciplinary and critical paths—and in all other moments related to the presentation and to the product advertising. Finally, decision makers at each level must become fully aware of the nature and functioning of AI systems to draw up appropriate rules for their use.

6. Conclusions

In this paper, we presented an ethical approach for defining a set of practical obligations and recommendations for the development of systems based on AI techniques. The main point of novelty of this work consists in an integrated view that takes into consideration both human rights and ethical foundations to reap the benefits of the two approaches. Moreover, it incorporates a specific focus on the Italian Constitution and proposes innovative research, among other topics, on redress and governance.

Although we are aware that general research on the ethical foundations of socially beneficial technological innovation are not enough to assure what they are supposed to achieve, we also believe that such efforts are necessary to spread awareness and encourage the establishment of good practices throughout all the relevant social areas—from industry to policy making, from distribution to education, and so on. We envision that researchers, policymakers and industry will take into account these high level recommendations in the most pressing effort of developing ethical and socially beneficial technologies.

Future research is, therefore, needed to assess the extent to which our recommendations can be translated into practice, including by means of standards, design principles or governance frameworks. We envision that our work will support decision-making processes and public debate aimed at the development of ethically oriented and trustworthy AI technologies. The ongoing discussion on the proposed AI Act in the European Union and the contextual debate on the Italian AI strategy call for a joint discussion on their compatibility, to be framed also in the context of the EU Green Deal and the Italian National Plan for Resistance and Resilience (PNRR). In particular, the AI Act proposal represents a unique opportunity to govern AI systems in a way that ensures that their development is placed within the context of fundamental rights. While this paper does not aim to suggest modifications on the Proposal, it contributes

[6] A prominent example of free online courses on AI is https://www.elementsofai.com/—Last accessed on January, 28th 2022.

to incrementing the hermeneutic pillars that will support its interpretation following its adoption, in particular the fundamental rights and freedoms that shall be respected throughout the whole AI systems' lifecycle. In the proposed Act, much is left to mixed forms of governance, with standardization entities playing a crucial role. This legislative strategy—known as the New Legislative Framework—calls for the contribution of private entities to co-regulate the design, development, and deployment of AI systems. Therefore, our recommendations also contribute to standardization initiatives by identifying principle-based requirements that should be implemented also by technical means. The promotion of fundamental rights and human-centric approaches, such as the ones discussed in this paper, are necessary to ensure that the development and the deployment of AI systems ultimately promote the 'common good' for the society that these regulatory interventions are shaping.

Acknowledgements All authors contributed equally to this work. This research was partially supported by Fondazione Leonardo - Civiltà delle Macchine (Italy).

References

Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, Bonnefon JF, Rahwan I (2018) The moral machine experiment. Nature 563(7729):59–64. https://doi.org/10.1038/s41586-018-0637-6

Broekens J, Heerink M, Rosendal H et al (2009) Assistive social robots in elderly care: a review. Gerontechnology 8(2):94–103. https://doi.org/10.4017/gt.2009.08.02.002.00

Capucha L, Nunes N, Calado A (2020) Artificial intelligence as a resilient tool for fighting inequalities in the covid-19 crisis. Eur J Eng Form Sci 4(2):10–19

Chalmers D, Davies G, Monti G (2019) European union law. Cambridge University Press, Cambridge

Cinelli M, De Francisci MG, Galeazzi A, Quattrociocchi W, Starnini M (2021) The echo chamber effect on social media. Proc Natl Acad Sci USA 118:1–8. https://doi.org/10.1073/pnas.2023301118

Cowls J, King T, Taddeo M, Floridi L (2020) How to design AI for social good: seven essential factors. Sci Eng Ethics 26:1771–1796. https://doi.org/10.1007/s11948-020-00213-5

Eubanks V (2018) Automating inequality How high-tech tools profile, police, and punish the poor. St. Martin's Press, London

Floridi L (2014) The fourth revolution. How the infosphere is reshaping human reality. Oxford University Press, Oxford

Floridi L (2015) The online manifesto: being human in a hyperconnected era. Springer, Cham

Floridi L (2019) Establishing the rules for building trustworthy AI. Nat Mach Intell 1(6):261–262. https://doi.org/10.1038/s42256-019-0055-y

Floridi L, Sanders JW (2004) On the morality of artificial agents. Mind Mach 14(3):349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Luetge C, Madelin R, Pagallo U, Rossi F, Schafer B, Valcke P, Vayena E (2018) AI4people—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Mind Mach 28(4):689–707. https://doi.org/10.1007/s11023-018-9482-5

Frey CB, Osborne MA (2017) The future of employment: how susceptible are jobs to computerisation? Technol Forecast Soc Change 114(1):254–280. https://doi.org/10.1016/j.techfore.2016.08.019

Goralski MA, Tan TK (2020) Artificial intelligence and sustainable development. Int J Manag Educ 18(1):1–9. https://doi.org/10.1016/j.ijme.2019.100330

Haner J, Garcia D (2019) The artificial intelligence arms race: Trends and world leaders in autonomous weapons development. Global Pol 10:331–337. https://doi.org/10.1111/1758-5899.12713

Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nat Mach Intell 1(9):389–399. https://doi.org/10.1038/s42256-019-0088-2

Kandlhofer M, Steinbauer G (2021) AI K–12 education service. Künstl Intell 35:125–126. https://doi.org/10.1007/s13218-021-00715-9

Kharbat FF, Alshawabkeh A, Woolsey ML (2021) Identifying gaps in using artificial intelligence to support students with intellectual disabilities from education and health perspectives. Aslib J Inf Manag 73(1):101–128. https://doi.org/10.1108/AJIM-02-2020-0054/full/html

Korinek A, Stiglitz JE (2019) Artificial intelligence and its implications for income distribution and unemployment. In: Agrawal A, Gans J, Goldfarb A (eds) The economics of artificial intelligence. Communications in computer and information science. University of Chicago Press, Chicago, pp 349–390

Lazer DMJ, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, Metzger MJ, Nyhan B, Pennycook G, Rothschild D, Schudson M, Sloman SA, Sunstein CR, Thorson EA, Watts DJ, Zittrain JL (2018) The science of fake news. Science 359(6380):1094–1096. https://doi.org/10.1126/science.aao2998

Lucivero F (2020) Big data, big waste? A reflection on the environmental sustainability of big data initiatives. Sci Eng Ethics 26:1009–1030. https://doi.org/10.1007/s11948-019-00171-7

Lynskey O (2014) Deconstructing data protection: the added-value of a right to data protection in the EU legal order. Int Comp Law Q 63(3):569–597. https://doi.org/10.1017/S0020589314000244

McGregor S (2021) Preventing repeated real world AI failures by cataloging incidents: the AI incident database. Proc AAAI Conf Artif Intell 35:15458–15463

Mittelstadt B, Allo P, Taddeo M, Wachter S, Floridi L (2016) The ethics of algorithms: mapping the debate. Big Data Soc. https://doi.org/10.1177/2053951716679679

Neto JSDO, Silva ALM, Nakano F, Perez-Alcazar JJ, Kofuji ST (2019) When wearable computing meets smart cities: assistive technology empowering persons with disabilities. Smart cities and smart spaces: concepts, methodologies, tools, and applications. IGI Global, Herschey, pp 1356–1376

Nishant R, Kennedy M, Corbetta J (2020) Artificial intelligence for sustainability: challenges, opportunities, and a research agenda. Int J Inf Manag 53:1–13. https://doi.org/10.1016/j.ijinfomgt.2020.102104

O'Neill C (2016) Weapons of math destruction. How big data increases inequality and threatens democracy. Penguin Books, London

Pagallo U (2020) Algoritmi e conoscibilità. Rivista Di Filosofia Del Diritto 9(1):93–106. https://doi.org/10.4477/97022

Palmirani M (2020) Big data e conoscenza. Rivista Di Filosofia Del Diritto 9(1):73–92. https://doi.org/10.4477/97021

Pasquale F (2015) The black box society. Harvard University Press, Harvard

Paul JA, Baker HM, Cochran JD (2012) Effect of online social networking on student academic performance. Comput Hum Behav 28(6):2117–2127. https://doi.org/10.1016/j.chb.2012.06.016

Quintarelli S (2019) Capitalismo immateriale: Le tecnologie digitali e il nuovo conflitto sociale. Bollati Boringhieri, Torino

Rachels J (1975) Why privacy is important. Philos Public Aff 4(4):323–333

Rodotà S (2015) Il diritto di avere diritti. Gius. Laterza & Figli Spa, Rome

Sirbu A, Giannotti F, Pedreschi D, Kertesz J (2019a) Public opinion and algorithmic bias. ERCIM News 116:15–16

Sirbu A, Pedreschi D, Giannotti F, Kertesz J (2019b) Algorithmic bias amplifies opinion fragmentation and polarization: a bounded confidence model. PLoS ONE 14(3):e0213246. https://doi.org/10.1371/journal.pone.0213246

Taddeo M, Floridi L (2018a) How AI can be a force for good. Science 361(6404):751–752. https://doi.org/10.1126/science.aat5991

Taddeo M, Floridi L (2018b) Regulate artificial intelligence to avert cyber arms race. Nature 556:296–298. https://doi.org/10.1038/d41586-018-04602-6

Vaccari C, Chadwick A (2020) Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Social Media 6(1):1–13. https://doi.org/10.1177/2056305120903408

Vedder A (2019) Safety, security and ethics. In: Vedder A, Schroers J, Ducuing C, Valcke P (eds) Security and law. Legal and ethical aspects of public security, cyber security and critical infrastructure security. Intersentia, Cambridge, pp 11–26

Villani D, Cipresso P, Gaggioli A, Riva G (2016) Integrating technology in positive psychology practice. IGI Global, Hershey

Vinuesa R, Azizpour H, Leite I, Balaam M, Dignum V, Domisch S, Felländer A, Langhans SD, Tegmark M, Fuso Nerini F (2020) The role of artificial intelligence in achieving the sustainable development goals. Nat Commun. https://doi.org/10.1038/s41467-019-14108-y

Warren SD, Brandeis LD (1890) The right to privacy. Harv Law Rev 4:193–220

Yan H, Jiang Y, Zheng J, Peng C, Li Q (2006) A multilayer perceptron-based medical decision support system for heart disease diagnosis. Expert Syst Appl 30(2):272–281. https://doi.org/10.1016/j.eswa.2005.07.022

Agenzia per l'Italia Digitale (2018) Libro bianco sull'intelligenza artificiale al servizio del cittadino. https://ia.italia.it/assets/librobianco.pdf

Alonso C, Kothari S, Rehman S (2020) Could artificial intelligence widen the gap between rich and poor nations? World Economic Forum. https://blogs.imf.org/2020/12/02/how-artificial-intelligence-could-widen-the-gap-between-rich-and-poor-nations/. Accessed 28 Jan 2022

Amershi S, Weld D, Vorvoreanu M et al (2019) Guidelines for human-AI interaction. In: Proceedings of the 2019 chi conference on human factors in computing systems. Association for Computing Machinery, New York, pp 1–13. https://doi.org/10.1145/3290605.3300233

Arntz M, Gregory T, Zierahn U (2016) The risk of automation for jobs in OECD countries: a comparative analysis. OECD social, employment and migration working papers 189:1–34. https://doi.org/10.1787/5jlz9h56dvq7-en

Canca C (2019) AI and global governance: human rights and AI ethics. Why ethics cannot be replaced by the UDHR. https://cpr.unu.edu/publications/articles/ai-global-governance-human-rights-and-ai-ethics-why-ethics-cannot-be-replaced-by-the-udhr.html. Accessed 28 Jan 2022

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) ImageNet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, p 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Drigas AS, Ioannidou RE (2013) A review on artificial intelligence in special education. In: Lytras MD, Ruan D, Tennyson RD, Ordonez De Pablos P, García Peñalvo FJ, Rusu L (eds) Information systems, e-learning, and knowledge management research. WSKS 2011. Communications in computer and information science, vol 278. Springer, Berlin, p 385–391. https://doi.org/10.1007/978-3-642-35879-1_46

European Commission for the Efficiency of Justice (CEPEJ) (2018) European ethical charter on the use of artificial intelligence in judicial systems and their environment. https://www.coe.int/en/web/cepej/cepej-european-ethical-charter-on-the-use-of-artificial-intelligence-ai-in-judicial-systems-and-their-environment. Accessed 28 January 2022

European Commission High Level Expert Group on Artificial Intelligence (2018) Ethics guidelines for trustworthy AI. https://ec.europa.eu/digital-single-market/en/news/ethicsguidelines-trustworthy-ai. Accessed 28 Jan 2022

Future of Life Institute (2017) Asilomar AI principles. https://futureoflife.org/2017/08/11/ai-principles/. Accessed 28 Jan 2022

House of Lords Select Committee (2018) AI in the UK: ready, willing and able. https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf

IEEE (2017) Ethically aligned design: a vision for prioritizing human well-being with autonomous and intelligent systems, version 2. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf

Loreggia A, Mattei N, Quintarelli S (2020) Artificial intelligence research for fighting political polarisation: a research agenda. In: Proceedings of the first international forum on digital and democracy. Towards a sustainable evolution 2020, Venice, Italy, December 10–11, 2020, volume 2781 of CEUR workshop proceedings, p 24–33

Partnership on AI (2016) Tenets. https://partnershiponai.org/about/#tenets. Accessed 28 Jan 2022

Pham QC, Madhavan R, Righetti L, Smart W, Chatila R (2018) The impact of robotics and automation on working conditions and employment. In: IEEE robotics and automation magazine June, p 126–128. https://ieeexplore.ieee.org/document/8385401

Quintarelli S, Corea F, Fossa F, Loreggia A, Sapienza S (2019) An ethical perspective on artificial intelligence: principles, rights and recommendations. p 159–177

Rossi A, Ducato R, Haapio H, Passera S (2019) Legal design patterns: towards a new language for legal information design. In: 22nd international legal informatics symposium IRIS 2019. https://cris.unibo.it/retrieve/handle/11585/673870/438421/legal_design_patterns_new_language_legal_info_design.pdf

SAE International (2018) J3016: taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. https://www.sae.org/standards/content/j3016_201806/

Tsiao JCs, Chao DY, Tong PP (2007) Natural-language voice activated personal assistant, May 8 2007. US Patent 7,216,080.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.