UNIO - *EU Law Journal*. Vol. 10, No. 1, May 2024, pp. 30-41.
®2024 Centre of Studies in European Union Law
School of Law – University of Minho

Building on the EU's unique strategy for Artificial Intelligence (AI): can an ethical foundation be successfully integrated into its design and deployment?

Maria Inês Costa[*]

ABSTRACT: In this article, the Author investigates the ethical principles espoused by the European Union's (EU) policies on Artificial Intelligence (AI) and discuss their effectiveness in providing an ethical foundation for the entire process of developing and deploying AI in Europe. To this end, the scope of the term "ethics" in the regulatory process is analysed, and the extent to which ambiguity about its meaning can contribute to inadequate policies, is considered. The Author also addresses the criticisms and suggestions that have been made regarding the use of certain concepts, such as 'trustworthiness' and 'human-centric', which are so often used in European AI guidelines. The aim of the article is to explore the ways in which ethical values have been applied and identify potential areas of improvement to further safeguard and advance the rule of law, human rights, and democracy in the EU. This quest proves to be relevant amidst the various challenges that arise from emerging technologies like AI, which are of an unknown magnitude to our contemporary democracies, constituting a complete paradigm shift. In exploring the issues set out here, the Author navigates some of the new developments that the EU's Artificial Intelligence Act (AI Act) introduces, as well as the pressing criticisms levelled at it in the context of the overarching theme of this article.

KEYWORDS: Ethical alignment – trustworthy artificial intelligence – human-centric – paradigm shift – Artificial Intelligence Act.

[*] Master in Human Rights and PhD candidate in Law at the University of Minho Law School. FCT scholarship holder - UI/BD/154522/2023.

1. The ethical and human-centric stance on AI: brief overview

The EU's ethical spirit already exists, in part, "*in terms of all the opening clauses […] in EU law, through the EU's values, including human dignity, the EU's human rights, as well as the 'spiritual and moral heritage'* [...]", as Markus Frischhut asserts.[1] The author goes on to highlight that the *Van Gend en Loos* judgment[2] established the EU as a «"*new legal order" with its unique spirit and legal system. Recognising this spirit comes with the understanding that it evolves over time and must adapt to the circumstances at hand. Hence, the EU has developed and established its own values and ethos over the course of its history*.»[3] In this sense, Mariachiara Tallacchini highlights how the term ethics embodies two distinct meanings in the EU: i) philosophical ethics, as a topic of discussion in the field of philosophy, and ii) "*ethics*" (in commas for emphasis), as a soft regulatory instrument that has flourished and contributes to guiding the legislative process within the context of European integration.[4]

It is precisely the deployment of this regulatory instrument that has played a pivotal role in shaping the applicable rules for Artificial Intelligence (AI) systems: [5] as stated in a European Parliament's (EP) document on AI, "*the EU can be considered a front-runner with regard to establishing a framework on ethical rules for AI.*"[6] Moreover, Inga Ulnicane elaborates on the EU's stance on AI policy: "*the goal is to ensure an appropriate ethical and legal framework that is based on the EU's values, in line with the Charter for Fundamental Rights. This includes guidance on regulation in […] areas of safety and liability, cooperation of stakeholders and development of AI ethics guidelines. An overarching idea highlighted in the strategy is that Europe should champion «an approach to AI that benefits people and society as a whole» and «place the power of AI at the service of human progress*".[7] The EU's spiritual and moral heritage is reflected in the way it approaches AI and the corresponding challenges it has faced over time, advocating a framework that protects fundamental rights and promotes democracy, namely through a strategy that empowers subjects and preserves human agency.

[1] Markus Frischhut, *The Ethical Spirit of EU Law* (Cham, Switzerland: Springer, 2019), 140.

[2] "The Court points out that the Community is a new legal order of international law in benefit of the states that had limited their sovereign. Furthermore, the Court argued that Community law imposes obligations for Member States and for nationals, and wanted to confer rights for the parts, arisen by the Treaty." See José Ricardo Sousa, "Summary of Van Gend en Loos – Case 26/62", *The Official Blog of UNIO - EU Law Journal, Thinking and Debating Europe*, March 3, 2016, https://officialblogofunio.com/2016/03/03/summary-of-van-gend-en-loos-case-2662/.

[3] Markus Frischhut, *The Ethical Spirit of EU Law*, 141.

[4] Mariachiara Tallacchini, "Governing by Values. EU Ethics: Soft Tool, Hard Effects", *Minerva* 47 (2009): 282. Doi: 10.1007/s11024-009-9127-1.

[5] See Artificial Intelligence Act, European Parliament legislative resolution of 13 March 2024 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), https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf, Article 3(1): "*'AI system' means a machine-based system designed to operate with varying levels of autonomy, that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.*"

[6] European Parliament, "EU guidelines on ethics in artificial intelligence: Context and implementation", EPRS | European Parliamentary Research Service, PE 640.163, September 2019, 2, EU guidelines on ethics in artificial intelligence: Context and implementation (europa.eu).

[7] Inga Ulnicane, "Artificial intelligence in the European Union – Policy, ethics and regulation", in *The Routledge Handbook of European Integrations*, ed. Thomas Hoerber, Gabriel Weber and Ignazio Cabras (London and New York: Routledge, 2022), 259.

Despite the emphasis this strategy has on the European arena, a report published in 2020 by the Council of Europe's Committee on Political Affairs and Democracy, entitled "*The need for democratic governance of artificial intelligence*", highlights, among other things, that there has been a trend in recent years towards the de-politisation of decision-making, with some individuals preferring AI to politicians when it comes to making political decisions. This was revealed in a 2019 survey on Europeans' attitudes towards technology, which found that a quarter of respondents preferred AI decision-making, despite it being based on statistical correlations rather than causal relationships. This shift in attitudes therefore reflects a growing distrust of governments and politicians and calls into question the Western model of representative democracy. The use of opaque and unaccountable algorithms in decision-making poses a serious threat to democratic values such as transparency, accountability and equality.[8]

Thus, while the key ideas outlined in this section are laudable and are included in the EU's instruments to regulate AI, it is clear that the emergence of this technology poses serious threats to the fundamental democratic and ethical values on which the EU was founded, and there is a need to better understand these challenges in order to find appropriate solutions. The following sections explore this in more depth.

2. AI in constitutional democracies: a paradigm shift

We live at a time where the risks that emerging technologies pose to us are of an unknown magnitude, and deciphering solutions is difficult for a myriad of reasons. As Ballaguer Callejón points out, the digital world – which is becoming an increasingly important part of our daily reality – is subject to rules in the production of which the state has virtually no role and which are not in line with constitutional principles and values, thus contributing to the fragmentation of the public space. Constitutionalism was forged in an analogue world, and the digital world – permeated by the omnipresence of AI, we should add – has changed the scope of its application.[9]

Moreover, as Federico Bueno da Mata stresses, when discussing the automation of certain jobs through the use of robots or the integration of AI in tasks with a degree of automation, we not only promote the development and commercialisation of robotics but also recognise a reality that our legal framework must promptly address. This brings with it a range of legal challenges to various fields of knowledge that require urgent attention.[10]

Indeed, AI is a disruptive technology that challenges conventional methods of explaining and organising the world, as it generates patterns and predictions that people struggle to understand and articulate.[11] The advancement of technology

[8] Council of Europe, "The need for democratic governance of artificial intelligence", Committee on Political Affairs and Democracy. Rapporteur: Ms Deborah Bergamini, Italy, Group of the European People's Party. Doc. 15150, 24 September 2020, 12.

[9] Francisco Balaguer, "La Constituición del algoritmo", *Estudos* 9, vol. 1 (2021): 24-26 and 30-31.

[10] Federico Bueno da Mata, "Retos jurídicos de la robótica: especial referencia al Derecho Procesal", in *Inteligência Artificial e Robótica – desafios para o Direito do século XXI*, ed. Sónia Moreira and Pedro Miguel Freitas (Coimbra: Gestlegal, 2022), 11.

[11] See Alessandra Silveira and Maria Inês Costa, "Regulating Artificial Intelligence (AI): on the civilisational choice we are all making", *The Official Blog of UNIO – EU Law Journal, Thinking and Debating Europe*, Editorial of July 2023, July 7, 2023, https://officialblogofunio.com/2023/07/17/editorial-of-july-2023/.

has resulted in substantial transformations in the human-technology relationship, and this is due to the significant increase in the level of intelligence exhibited by computers, which has surpassed the previous conception of them being tools solely used for calculations or classifications. Their present ability to perform tasks that are comparable to those of autonomous human actions has led to a shift in our experience of technology. Thus, the impact of AI cannot merely be measured through conventional quantitative methods, as its effects extend beyond these boundaries. The proliferation of this technology has led to an era in which our interaction with technology is intertwined and subjective,[12] and this raises questions about our perception of reality, the role of humans within it, and the future of democratic societies and free will.

In this context, as Amanda Lagerkvist puts it, "*we may ask […] whether Big Data, AI and machine learning of the present age, with their technocratic, entrepreneurial and capitalistic ethos, will further hamper […] the prospects for realizing ourselves through projects of our will. Or will they even relieve humans of the responsibility they have for their lives, for each other, and for the planet?*".[13] These are questions that reflect the concerns outlined above, while also embracing the beneficial possibilities of AI in our world.

At the same time, the innovation that AI can bring, and its benefits do not make the great dangers disappear – the risks are growing in number and magnitude and make us question the extent to which we should sacrifice security for the sake of innovation.

2.1 Risks and challenges posed to the future of democracy

In fact, the concept of risk has been extensively developed by the German sociologist Ulrich Beck, who considers today's society to be a "*risk society*", a term which is also part of the title of his famous work, originally published in 1986 – *Risikogesellschaft: Auf dem Weg in eine andere Moderne*.[14] Here we encounter the notion that attention has turned from the development and utilisation of technologies across different fields to the political and economic management of the risks pertaining to these technologies. This implies the identification, treatment, recognition, prevention, or concealment of hazards in specific contexts of relevance and vulnerability.

Hence, the concept of security is frequently strengthened in response to growing threats and potential damage, often requiring interventions in technological and economic progress to maintain the confidence of citizens. Are risks not already inherent to the period of industrial society? In response to this query, Beck contends that modernity did not invent risks: in the past, individuals who embarked on discovering new territories accepted personal risks. However, current worldwide threats differ from the previous perception of "*risk*" as being bold and daring. Instead, it now represents a destructive danger to life at large.[15]

Indeed, AI systems can produce unfair and discriminatory outcomes that undermine democratic processes and negatively impact vulnerable communities, so

[12] Danilo Cesar Maganhoto Doneda, *et al.*, "Considerações iniciais sobre inteligência artificial, ética e autonomia pessoal", *Pensar - Revista de Ciências Jurídicas*, Fortaleza, vol. 23, no. 4 (2018): 2. Doi: 10.5020/2317-2150.2018.8257.

[13] Amanda Lagerkvist, «Digital limit situations: anticipatory media beyond 'the new AI era'», *Journal of Digital Social Research*, vol. 2, no. 3 (2020): 23.

[14] Ulrich Beck, *Risikogesellschaft: Auf dem Weg in eine andere Moderne* (Frankfurt am Main: Suhrkamp, 1986).

[15] Ulrich Beck, *Risk Society* (Sage Publications: 1992), 21.

careful consideration and oversight of AI decision-making is necessary to preserve democratic legitimacy.[16] In this respect, there is currently a strong discussion regarding the need to effectively implement ethical principles with respect to AI systems, which underlines the numerous proposals calling for the incorporation of ethical considerations into the regulation of AI.[17]

In the subsequent section, our intention is to examine and provide insight on how ethical values have been defined and established. Moreover, we identify criticisms and suggestions for improvement to further safeguard and promote the rule of law, human rights and democracy in the EU.

3. AI "made in Europe": analysing the core elements and recent developments

The EU has on numerous occasions underlined its commitment to providing guidance on AI based on ethical standards and respect for the values set out in the Charter of Fundamental Rights (CFREU). This has been the case since 2017, when the European Parliament (EP) launched its Resolution on Civil Law Rules on Robotics.[18] In regulating AI, the EU has made strides, by establishing guidelines, directives, and more recently, the Artificial Intelligence Act (AIA). As suggested by Erik Brattberg *et al.*, European leaders have stressed that AI is a priority, with Europe aiming to become a global leader in this domain. The emphasis is placed on AI that is "*made in Europe*", ethical, human-centric and that aligns with human rights and democratic values.

In the same vein, the European Commission aims to leverage its regulatory and market power to gain a competitive edge in the AI industry under the banner of "*trustworthy AI*",[19] in line with the *ethos* of this shared space. When it comes to AI "*made in Europe*", the Commission argues that the EU's approach to AI is unique: "*while actions are geared towards developing technology that is competitive and makes the most of the opportunities offered by AI, this technology should also be* ***ethical*** *and* ***secure****.*"[20]

On 13 March 2024, the European Parliament formally approved the AI Act, on what was labelled, according to *Euronews*, as a historic day by Brando Benifei, an Italian MEP who co-led the Parliament's AI Act, adding that this is the "*first regulation in the world which puts a clear path for a safe and human centric development of AI*".[21] Indeed, Recital 8 of the Regulation lays out the support for the objective of a "*European human-centric approach to AI*", which "*ensures the protection of ethical principles, as specifically requested by the European Parliament.*"[22]

[16] Council of Europe, "The need for democratic governance...", 12.

[17] Council of Europe, "The need for democratic governance...".

[18] On the evolution of approaches to AI in the EU, see Inga Ulnicane, "Artificial intelligence in the European Union – Policy, ethics and regulation", in *The Routledge Handbook of European Integrations*, ed. Thomas Hoerber, Gabriel Weber and Ignazio Cabras (London and New York: Routledge, 2022).

[19] Erik Brattberg, Raluca Csernatoni, and Venesa Rugova, *Europe and AI: Leading, Lagging Behind, or Carving Its Own Way?* (Carnegie Endowment for International Peace, 2020), 1.

[20] European Commission, "Questions and Answers: coordinated plan for Artificial Intelligence «made in Europe»", 7 December 2018, Brussels, available at: https://ec.europa.eu/commission/presscorner/detail/en/memo_18_6690. Our bold.

[21] Cynthia Kroet, "Lawmakers approve AI Act with overwhelming majority", *Euronews*, 13 March 2024, updated on 14 March 2023, https://www.euronews.com/my-europe/2024/03/13/lawmakers-approve-ai-act-with-overwhelming-majority.

[22] See Artificial Intelligence Act, European Parliament legislative resolution of 13 March 2024 on

The provisions of this Regulation entail harmonisation for the placing on the market, putting into service and use of AI systems, prohibition of certain practices and specific requirements for high-risk AI systems as well as for operators of such systems. Moreover, it is stated that "*diversity, non-discrimination and fairness means that AI systems are developed and used in a way that includes diverse actors and promotes equal access, gender equality and cultural diversity, while avoiding discriminatory impacts and unfair biases that are prohibited by Union or national law […]*". Thus, *"the application of those principles* ***should be translated, when possible, in the design and use of AI models*** *(…).*"

Although the legally binding risk-based approach[23] can contribute to creating more ethically sound AI technologies, as it establishes transparency obligations and requirements for the operators of these systems, as well as penalties in case of non-compliance with these mandatory commitments,[24] it seems that most efforts to develop ethical AI are still non-binding and require further debate.[25] In this sense, it is important to understand what ethics we are covering in practice and also to recognise that if this field remains mainly in the realm of non-binding legislation, it could lose its momentum and overall effectiveness in the long term.

3.1 Defining concepts: what kind of 'ethics'?

What is, in fact, trustworthy AI? Mona Simion and Christopher Kelp stress that "*policy makers and AI developers around the world have invested millions to answer this question*", and the reason why this is so relevant is because "*societies will only ever be able to achieve the full potential of AI if trust can be established in its development, deployment, and use.*"[26] However, various scholars, not just from the field of law but also from

the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act). See also European Parliament resolution of 20 October 2020 with recommendations to the Commission on a framework of ethical aspects of artificial intelligence, robotics and related technologies, 2020/2012(INL). This work was drafted after the AI was approved by the European Parliament, but before its formal approval by the Council and publication in the Official Journal. Therefore, there may be slight differences between the AI Act as cited here and the version published in the Official Journal resulting from linguistic revision.

[23] See Iakovina Kindylidi and Tiago Cabral, "Proposal for a Regulation on a European Approach for Artificial Intelligence: An Overview", *Whatnext.law*, May 5, 2021, https://whatnext.law/en/2021/05/05/proposal-for-a-regulation-on-a-european-approach-for-artificial-intelligence-an-overview-en/.

[24] See Artificial Intelligence Act, European Parliament legislative resolution of 13 March 2024, Chapter XII, Article 99. It provides that Member States shall implement rules on penalties and other enforcement measures, which may also include warnings and non-monetary measures, applicable to breaches of the Regulation by operators, and must take all necessary measures to ensure that they are properly and effectively enforced, taking into account the guidelines issued by the Commission pursuant to Article 96. What is more, "*the penalties provided for shall be* ***effective, proportionate and dissuasive*** *[...]*".

[25] "As one can observe, common elements are evident in various EU initiatives concerning the development and regulation of AI systems, often employing keywords or key expressions such as trust, safety, traceability, acceptability, risk assessment, fairness, non-discrimination, and so on. While all of regulation efforts are to be commended, several authors point to the need to delve deeper into what these terms really mean, the true impact of these new rules in a world that is undergoing a real paradigm shift, and the need to evaluate the emergent approaches critically and continuously to the use of AI in contemporary democracies." See Maria Inês Costa, EU's policies to AI: are there blindspots regarding accountability and democratic governance?, *The Official Blog of UNIO – Thinking and Debating Europe*, October 27, 2023, https://officialblogofunio.com/2023/10/27/eus-policies-to-ai-are-there-blindspots-regarding-accountability-and-democratic-governance/#_ftn5.

[26] Mona Simion and Christoph Kelp, "Trustworthy artificial intelligence", *Asian Journal of Philosophy*, Springer 2:8 (2023): 1.

philosophy, sociology, and the technical sphere itself, have reflected upon and offered criticism regarding the ethical standards that we find referred to in all those proposed instruments. This merits discussion, as some argue that the field of ethics does not sufficiently support the secure advancement of AI for humans, and others contend that there is a lack of comprehension regarding ethical codes and how they can – or if they can –, be integrated into the development of AI systems.

For instance, Brent Mittelstadt argues that employing normative ideas such as "*fairness*" and "*dignity*" is vague and excessively abstract. This results in ambiguity in comprehension, as the mentioned concepts can have contrasting meanings depending on one's beliefs – they are "*essentially contested concepts*". Hence, because of this ambiguity, appealing to those concepts often obscures underlying disagreements on political and ethical values, and different interpretations can lead to distinct approaches to AI system development and implementation. Also, it should be noted that even if a consensus is reached on ethical principles for AI in theory, this does not mean that agreement on practical implications will be reached – these are two different realms.[27] As the author suggests, during the implementation phase, developers may encounter conflicting moral principles and frameworks that lead to genuine ethical dilemmas beyond the scope of principlism.[28]

Thilo Hagendorf reminds us that "*in AI ethics, technical artefacts are primarily seen as isolated entities that can be optimized by experts so as to find technical solutions for technical problems. What is often lacking is a consideration of the wider contexts and the comprehensive relationship networks in which technical systems are embedded.*"[29] We believe that the networks in which technical systems are integrated involve the connections between individuals in societies, their relationships with institutions, and the principles that these embody, acknowledged by people who act according to a specific set of values prevalent in their respective societies. These interdependent networks are influential in the creation and advancement of everything. Thus, as nothing exists in isolation, it is crucial to establish connections between different fields of knowledge to formulate sound theories that can provide viable solutions to emerging issues, such as applying an ethical approach to the development of AI.

[27] Brent Mittelstadt, "Principles alone cannot guarantee ethical AI", *Nature Machine Inteliggence* vol. 1, no. 11 (2019): 503.

[28] *I.e.,* the framework that draws upon the bioethical framework pioneered by Tom Beauchamp and James Childress in 1979 and which is based on four core principles – autonomy, beneficence, nonmaleficence and justice. «For Beauchamp and Childress, the common morality is what they take to be a universal morality, one to which all morally serious persons are committed [...]. The content of the common morality is dictated by the primary objectives of morality, which include, for example, the amelioration of human misery. It encompasses certain rules of obligation (tell the truth, keep promises), and endorses certain standards of moral character, such as honesty and integrity [...] Importantly, this common morality is historicist, in that its authority is established historically, through the success of its related norms in advancing human flourishing across time and place. However, unlike many historicist accounts, the common morality is not relativist, as its norms are to be applied universally. [...] A fairly recent criticism comes from John McMillan, according to whom Beauchamp and Childress' approach stifles careful reflection about real issues. McMillan claims that principle-centered methods cannot lead to the formulation of what he calls "reasoned convictions about moral problems", and writes that the four principles approach hinders bringing moral reason to bear upon practical questions. [...]». On the principled-approach and its critique, see Jennifer Flynn, "Theory and Bioethics", *The Stanford Encyclopedia of Philosophy* (Winter 2022 Edition), Edward N. Zalta & Uri Nodelman (eds.), https://plato.stanford.edu/archives/win2022/entries/theory-bioethics/.

[29] Thilo Hagendorf, "The Ethics of AI Ethics: An Evaluation of Guidelines", *Minds and Machines* 30:99 (2020): 103.

Along these lines, Mark Coeckelbergh states that AI is not a separate entity, but dependent on other technologies and integrated into broader scientific practices. Thus, the ethical dimensions of AI should be linked to wider ethical issues around digital information, communication technologies, and computer ethics. The author insists AI is not just about technology, but also about how humans interact with and use it: one must understand human perceptions and experiences of AI, and its place within the broader socio-technical environment. Thus, AI must be viewed as part of a complex system involving technology and human society, and its ethical implications should be considered accordingly.[30]

In this sense, Bernd Carsten Stahl *et al.* present us with the perspective that ethical theories aid in understanding the reasons for ethical concerns and how to handle them. However, the current discourse on ethics and AI tends to disregard foundational ethical theories, and rather centres around principles, which are "*mid-level applicable concepts*" used to steer action. Moreover, although there is increasing advocacy for AI ethics principles, comparative studies demonstrate that the number of such principles is relatively low.[31]

In this context, it is important to observe Marc Anderson's criticism of the AI Act in its attempt to translate ethical tenets into legislation: "*there is no formal, or even tacit, acceptance of discourse theory to be found in discussions of the ethical consultations around AI which are said to precede the AI Act, nor such an acceptance of any other ethical theory*". Moreover, the author stresses that the development of AI law prioritises, first and foremost, democratic and majoritarian principles rather than law drawing on ethics – the current process aims to incorporate ethics to strengthen its position in law, as ethics alone is "*too weak*" for the desired outcome. This reflection attempts to show how ethical values may have been proclaimed in a more surface-level manner, not constituting the real foundation of the AI Act's legislation. Moreover, this also means that attempting to entrench ethical elements in legislation raises many challenges, not least pragmatic ones.[32]

These analyses show that the implementation of the famous motto "*ethics by design*" is a challenging undertaking, which requires us not only to deepen our knowledge from a philosophical perspective, but also to consider: i) what we are really framing in terms of ethics when regulating the development and deployment of AI; and ii) whether it is actually feasible to achieve the desired outcomes in the long run.

4. (In)effectiveness of "trustworthy" AI: a brief analysis

In the document prepared by the High-Level Expert Group on Artificial Intelligence (AI HLEG) – "*Ethics Guidelines for Trustworthy Artificial Intelligence (AI)*" –, the term "*trustworthy AI*" is defined as having three fundamental components -

[30] Mark Coeckelbergh, *AI Ethics* (Cambridge, Massachusetts; London, England: The MIT Press, 2020), 93.

[31] Bernd Carsten Stahl, *et al.*, "Exploring ethics and human rights in artificial intelligence – A Delphi study", *Technological Forecasting and Social Change,* vol. 191 (2023): 2. Doi: https://doi.org/10.1016/j.techfore.2023.122502.

[32] Marc Anderson, "Some ethical reflections on the EU AI Act", *IAIL 2022: 1st International Workshop on Imagining the AI Landscape After the AI Act*, 13 June (2022): 3, IAIL_paper5.pdf (ceur-ws.org). See also Thomas Powers and Jean-Gabriel Ganascia, "The ethics of the ethics of AI", in *The Oxford Handbook of AI*, ed. Markus D. Dubber, Frank Pasquale and Sunit Das (Oxford: Oxford University Press, 2020), 28.

it should be: i) lawful; ii) ethical; and iii) robust, both from a technical and social perspective. The ethical principles it must uphold are: (i) respect for human autonomy; (ii) prevention of harm; (iii) fairness; and (iv) explainability. Moreover, seven requirements must be met: (i) human agency and oversight; (ii) technical robustness and safety; (iii) privacy and data governance; (iv) transparency; (v) diversity; (vi) non-discrimination and fairness; and (vii) environmental and societal well-being and accountability. Finally, it should be noted that this framework is based on legally-protected fundamental rights in the EU, stemming from the Treaties, the CFREU, and international human rights law.

The AI Act provides for these standards to be highly recommended, with the aim of creating an AI that serves people, respects human dignity, and preserves personal autonomy.[33] To this end, scholars have written on how we can achieve this paradigm shift, where trust and explainability reign. For instance, Richa Singh *et al.*, drawing on the Guidelines mentioned above, argue that more openness and transparency in AI systems is needed in order to build trust. Since many of the current systems are opaque when it comes to their model lineage, the training of the data and performance details, minimum disclosure standards should be established.

In fact, the AI Act lays down obligations that aim to address this concern, for example by establishing in its Article 53 that providers of general-purpose AI models are obliged to "*draw up and keep up-to-date the technical documentation of the model, including its training and testing process and the results of its evaluation […]"* and shall *"make available information and documentation to providers of AI systems who intend to integrate the general-purpose AI model into their AI systems*", while still protecting intellectual property rights. This is one of the ways in which the Regulation aims to mandate transparency and explainability of AI technology,[34] fundamental qualities of the paradigm of *trustworthiness*. But as Cecilia Panigutti and others stress, when it comes to explainability and interpretability, even the concept of what constitutes a proper human understandable explanation is up for debate among the research community.[35]

For instance, we have Mona Simion and Christopher Kelp's account of two major problems with the trustworthy AI framework. According to the authors, it lacks explanatory adequacy, and what they mean by this is that we may consider some AI systems trustworthy, but without knowing the reasons why we think so: we are just relying on a case-by-case basis, and not on the nature of what makes something actually trustworthy.[36] Also, the framework under analysis suffers from

[33] See Artificial Intelligence Act, European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence, recital 27.

[34] See Artificial Intelligence Act, recital 27: "*[…] Transparency means that AI systems are developed and used in a way that allows appropriate traceability and explainability, while making humans aware that they communicate or interact with an AI system, as well as duly informing deployers of the capabilities and limitations of that AI system and affected persons about their rights. […]*"

[35] Cecilia Panigutti, Ronan Hamon, Isabelle Hupont, David Fernandez Llorca, Delia Fano Yela, Henrik Junklewitz, Salvatore Scalzo, Gabriele Mazzini, Ignacio Sanchez, Josep Soler Garrido, and Emilia Gomez, "The role of explainable AI in the context of the AI Act", *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*, June 12–15, Chicago, IL, USA. ACM, New York, NY, USA (2023): 1140, https://doi.org/10. 1145/3593013.3594069.

[36] "*[…] say that your preferred list of trustworthiness-making properties seems impeccably extensionally adequate — in that it seems to infallibly predict an AI is trustworthy when it is, and conversely, that it is not to be trusted when it is not. The question as to why your theory got it right remains unanswered: what is the trustworthy-making*

the tendency to misuse the term "*trustworthiness*", and to mistake it for mere reliance. Hence, something trustworthy is/tends to be universally reliable, but something which is reliable is not necessarily trustworthy. Indeed, these terms cannot be used interchangeably: reliability relates somewhat to consistency, for example, in a certain course of actions, but trustworthiness requires an "*extra factor*".[37]

Moreover, Hagendorf suggests that "*the practice of development, implementation and use of AI applications has very often little to do with the values and principles postulated by ethics", since "developers are neither systematically educated about ethical issues, nor are they empowered, for example by organizational structures, to raise ethical concerns. In business contexts, speed is everything in many cases and skipping ethical considerations is equivalent to the path of least resistance […]*".[38] This contributes to the ineffectiveness of ethics in AI, in the author's opinion.

In a broader sense, Luke Munn presents a highly critical view of the technological industry and explains why the ethics of AI appear futile considering the large corporations steering AI's growth. The author contends that unethical AI stems from an industrial culture that lacks concern for upholding ethical values, which is rooted in the broader scheme of the tech industry's lack of commitment to ethics throughout its history. There is a deficit of ethics education in the curricula of software engineering courses,[39] And if this is allowed to continue, the logical consequence is that ethics will never be a priority that the big tech companies want to address.

Furthermore, the pursuit of profit in the AI race can drastically alter the orientation of organisations and missions that aim to have a constructive impact on society using AI technology, prioritising security and the defence of fundamental rights, which is of paramount concern to liberal democracies. This is why the questions posed by Krasodomski and Buchser are so relevant: whether the so-called Brussels effect will be reflected in AI, and whether the new regulatory framework will have worldwide repercussions for the further advancement of this technology.[40]

4.1 Brief overview of the element of 'accountability'

Accountability is a very important component in the development and implementation of trustworthy AI, especially if we consider the EU's stance in this regard, as in the Guidelines discussed in this article. In fact, the AI Act repeatedly addresses this factor, as can be seen in the following passages:

> *"It is […] appropriate to classify as high-risk […] a number of AI systems intended to be used in the law enforcement context where accuracy, reliability and transparency is particularly*

underlying property that delivers one particular list rather than another? Why should we think, for instance, that explainability belongs on the list, while transparency does not? Conversely, if we think that, on closer inspection, we should include transparency as well, why is that so? Short of having an answer to this question, we run the risk that our list merely covers paradigmatic cases of trustworthy AIs, rather than the nature thereof". See Mona Simion and Christoph Kelp, "Trustworthy artificial intelligence", 2.

[37] "*[…] You rely on the weather not to suddenly drop by 20 degrees, leaving you shivering; you rely on your colleague at work to help you with your jammed printer, because they're just better at this stuff; you rely on the shop at the corner to still be there tomorrow when you need to buy milk. Trust, the thought goes, is a more precious and less ubiquitous commodity. For most philosophers, trust involves reliance "plus some extra factor" […]. The question as to what this extra factor might be has generated impressive amounts of literature in the ethics and epistemology of trust. In contrast, this distinction has been ignored in AI research.*" See Mona Simion and Christoph Kelp, "Trustworthy artificial intelligence", 2.

[38] Thilo Hagendorf, "The Ethics of AI Ethics: An Evaluation of Guidelines", 108.

[39] Luke Munn, "The uselessness of AI ethics", *AI and Ethics* (2023): 871. Doi: https://doi.org/10.1007/s43681-022-00209-w.

[40] Alex Krasodomski and Marjorie Buchser, "The EU's new AI Act could have global impact", *Chatham House*, 13 March 2024, https://www.chathamhouse.org/2024/03/eus-new-ai-act-could-have-global-impact.

> *important to avoid adverse impacts, retain public trust and ensure accountability and effective redress […]".*[41]
>
> *"European common data spaces established by the Commission and the facilitation of data sharing between businesses and with government in the public interest will be instrumental to provide trustful, accountable and non-discriminatory access to high quality data for the training, validation and testing of AI systems".*[42]
>
> *"[…] providers of general-purpose AI models with systemic risks should continuously assess and mitigate systemic risks, including for example by putting in place risk-management policies, such as accountability and governance processes, implementing post-market monitoring, taking appropriate measures along the entire model's lifecycle and cooperating with relevant actors along the AI value chain."*[43]

And what is accountability? In a broad sense, to be accountable refers to the fact of being responsible for what one does and to be able to give a satisfactory reason for it, or the degree to which this happens; it can also be defined as a situation in which someone is responsible for things that happen and can give a satisfactory reason for them.[44]

Drawing on Van de Poel and Sand's definition of accountability – "*prescriptive dimension as it presumes the ability and willingness to account for one's actions and to justify them to others*" –, Johan Rochel and Florian Evéquoz consider it a crucial component of our societies. The element of justification embedded in it reflects a fundamental human condition, namely the capability to reason, reflect, and explain, particularly in a community, where debate and deliberation are important.[45] It is more than just asserting a point of view, it is also about being able to make connections between concepts and to argue, thus exercising fundamental human faculties.

In the realm of AI, it "*relates to the expectation that designers, developers, and deployers will comply with standards and legislation to ensure the proper functioning of Ais during their lifecycle*", but "*Ais are neither mere artifacts nor traditional social systems: technological properties often make the outcome of Ais opaque and unpredictable, hindering the detection of causes and reasons for unintended outcomes.*"[46] The opacity of AI systems and their fundamentally different kind of intelligence applied to understanding reality can, in fact, create a very hostile environment for democracy to thrive. But there is more to it. As Laux, Wachter and Mittelstadt emphasise, "*well-placed trust in AI systems in the public sector requires institutions of public accountability*".[47] Whilst Recital 57 of the AI Act provides that the concept of '*AI literacy*' corresponds to "*skills, knowledge and understanding that allows* ***providers, deployers and affected persons*** *[…] to make an informed deployment of AI systems, as well as to gain awareness about the opportunities and risks of AI and possible harm*

[41] Artificial Intelligence Act, recital 59.

[42] Artificial Intelligence Act, recital 68.

[43] Artificial Intelligence Act, Recital 114.

[44] Definition of 'accountability' from the Cambridge Business English Dictionary © Cambridge University Press, accessed 23 November, 2023, https://dictionary.cambridge.org/dictionary/english/accountability?q=Accountability.

[45] Johan Rochel and Florian Evéquoz, "Getting into the engine room: a blueprint to investigate the shadowy steps of AI ethics", *AI & SOCIETY* 36 (2021): 613. Doi: https://doi.org/10.1007/s00146-020-01069-w.

[46] Claudio Novelli, Mariarosaria Taddeo, Luciano Floridi, "Accountability in artificial intelligence: what it is and how it works", *AI & Society* (2023): 5.

[47] Johann Laux, Sandra Wachter and Brent Mittelstadt, "Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk", *Regulation & Governance* 18 (2024): 18.

it can cause", there is a pressing knowledge gap between laypeople and AI developers/experts which makes genuine and informed participation a challenging task. Hence why suggestions of intermediary institutions as "*trust proxies*", motivated by the existence of knowledge asymmetries, have been put forward in the literature.[48]

Although we often uncritically use technology whose inherent processes we cannot grasp, the quality of explainability remains imperative. If we do not possess it, then democratic values such as personal autonomy and human dignity are jeopardised.[49]

Final Remarks

Throughout this article, our aim has been to encourage what we hope will be a fortuitous discussion in the context of AI regulation, which is often dominated by references to the need and priority of an ethical approach to its safe and beneficial development for humanity, the tenets of which are reiterated in the AI Act approved by the EP in March 2024.

It was observed, albeit briefly, that applying an ethical framework that has truly positive practical implications requires a deeper and broader knowledge of the field of ethics and a clear definition of what "*ethics*" we are trying to apply: whether we are simply deriving ideas from loose principles or trying to establish a framework based on foundational ethical theories.

In addition, we were able to observe how the regulation of AI must be framed within the broad context of the historical, social and cultural development of societies. The ability of the law to contribute to the regulation of this technology in accordance with the standards that the EU has so often affirmed must be strengthened by insights from other fields of knowledge, and only this multidisciplinary dialogue can lead to more robust frameworks.

Although we may seem to be left with more questions than answers, the truth is that AI, being a disruptive technology that calls into question many of our patterns of explanation, metamorphoses many of our ways of ordering and classifying the world, and the power it has in these onslaughts is of enormous proportions. In this sense, we need to ask questions and urgently address several of the loopholes that endanger democracies and the key values that underpin it. All things considered, it is still encouraging to witness the EU taking steps to legislate and regulate AI technologies, with all the challenges that it may present. After all, it is only possible to critique (and ultimately improve) what is already in motion.

[48] Johann Laux, Sandra Wachter and Brent Mittelstadt, "Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk", 19.

[49] See Alessandra Silveira and Maria Inês Costa, "Regulating Artificial Intelligence (AI): on the civilisational choice we are all making", *The Official Blog of UNIO - EU Law Journal, Thinking and Debating Europe*, Editorial of July 2023, July 7, 2023, https://officialblogofunio.com/2023/07/17/editorial-of-july-2023/.