The AI Act:
A realpolitik compromise and the need to look forward

Alessandro Mantelero

Abstract: First-generation technology regulation typically attempts to strike the right balance between rights protection and innovation. This tension is evident in the EU AI Act and in the way the risk management, the core element of any technology regulation, is framed. This chapter outlines the rationale behind the compromise solution adopted by the EU legislator to reconcile the protection of fundamental rights with the expected benefits of AI. It also discusses the decision to depart from a more holistic approach centred on the societal acceptability of AI, in terms of alignment with the values of the communities in which AI solutions are to be implemented. The chapter highlights the weakness of a primarily risk-based approach that does not place at the heart of the regulation the definition of key principles specifically tailored to the AI context and aimed at underpinning its development. Against this background, the role of fundamental rights in guiding the development of a human-centred AI in line with EU values is crucial. However, the implementation of the fundamental rights impact assessment in the AI Act is still underway. A more coherent framework is needed, combining the different assessments outlined in the AI Act, as well as a better definition of the scope and relevant criteria for the assessment. Finally, an appropriate model should be developed and made available AI providers and deployers, adopting a lean assessment design and combining expert-based evaluation and stakeholder/rightsholder participation.

A. Introduction

After a long debate on the impact of Artificial Intelligence (AI) on society, the European Union has decided to adopt the first legal instrument specif-

ically focused on this technology,[1] whose recent development, despite its many benefits, has raised several concerns in a variety of areas.

The EU was the first mover in this field within the global geo-political regulatory scenario, but this is not the only initiative to establish some rules for AI development and use. From Brazil[2] to the US,[3] many other lawmakers are outlining specific provisions for AI, and a number of charters providing key principles for AI development have been adopted by a variety of entities in recent years.[4] This is the typical scenario for a first generation of new technology regulation, as was the case with data protection in the late 1960s and early 1970s.

As discussed in the next section, a common problem for many first-generation technology regulations is finding the right balance between rights protection and innovation. This tension is also evident in the EU AI Act

1 More details on the AI Act and its approval process can be found here (all online resources referred to in the footnotes to this chapter have been consulted prior to 1 September 2023):
https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence?&at_campaign=20226-Digital&at_medium=Google_Ads&at_platform=Search&at_creation=RSA&at_goal=TR_G&at_advertiser=Webcomm&at_audience=artificial%20intelligence%20act&at_topic=Artificial_intelligence_Act&at_location=IT&gclid=EAIaIQobChMI9Z__pIqJgQMVsZ1oCR2jvw1ZEAAYASAAEgL6mfD_BwE.

2 See here the progress of the proposal: https://www25.senado.leg.br/web/atividade/materias/-/materia/157233?_gl=1*cqpafr*_ga*NzcyNDkwNDc3LjE2NTc2MzI1OTk.*_ga_CW3ZH25XMK*MTY4NDI0NDk5Mi4xMS4xLjE2ODQyNDU0NTcuMC4wLjA.# publicacoes. See also Belli, Luca, Yasmin Curzi, e Walter B. Gaspar. «AI regulation in Brazil: Advancements, flows, and need to learn from the data protection experience». *Computer Law & Security Review* 48 (1 April 2023): 105767. https://doi.org/10.1016/j.clsr.2022.105767.

3 For an overview of the different US regulatory initiatives, see also https://www.ravitdotan.com/us-ai-regulation.

4 For a more detailed analysis see Mantelero, Alessandro. Beyond Data: Human Rights, Ethical and Social Impact Assessment in AI. Vol. 36. Information Technology and Law Series. The Hague: T.M.C. Asser Press, 2022. https://doi.org/10.1007/978-94-6265-531-7 (Open Access), 93-101. See also Jobin A, Ienca M, Vayena E (2019) The Global Landscape of AI Ethics Guidelines. 1 Nature Machine Intelligence 389; Hagendorff T (2020) The Ethics of AI Ethics: An Evaluation of Guidelines. 30 Minds and Machines 99; Ienca M, Vayena E (2020) AI Ethics Guidelines: European and Global Perspectives. In: Council of Europe. Towards regulation of AI systems. Global perspectives on the development of a legal framework on Artificial Intelligence systems based on the Council of Europe's standards on
human rights, democracy and the rule of law. Council of Europe, Strasbourg, pp 38–60.

and in the way the core element of all technology regulation, namely risk management, is framed. In addition, despite a significant debate on the societal impact of AI – mainly from the point of view of AI ethics[5] – the AI Act is not (yet?) part of a holistic approach that combines legal and non-legal issues revolving around AI.[6]

Finally, the focus on the risk-based approach has paid less attention to the definition of key principles specifically tailored to the AI context, which are neither the core of the AI Act[7] nor adequately elaborated in the other regulatory initiatives and in the many guidelines on AI. As noted in the analysis of the proposed principles carried out in the fourth section, these principles are often vague, overlap with similar principles set out in in other regulations, without clarifying their relationship with them, and in any case require specific guidelines for their consistent and concrete implementation in AI design and development.

In the absence of a sound set of guiding principles underpinning the EU way to AI and with a regulatory focus primarily centred on risk management, the last section emphasises the key role played by human rights (fundamental rights within the EU context) in the development of a truly human-centred AI that embodies EU values. Given the structure of the AI Act and the pivotal role of the risk-based approach, the impact assessment on fundamental rights becomes crucial in order not to restrict this regulation to a mere safety and security perspective.

The issues briefly listed here and discussed in more detail in the following sections urge law-makers to make further efforts to define the core elements of a methodology for assessing the impact of AI on fundamental rights and to support its implementation, avoiding simplistic solutions

---

5 See fn. 4. See also European Data Protection Supervisor (2015b) Opinion 4/2015b. Towards a new digital ethics:
Data, dignity and technology, https://edps.europa.eu/sites/edp/files/publication/15-09-11_data_ethics_en.pdf; European Data Protection Supervisor, Ethics Advisory Group (2018) Towards a digital ethics.
https://edps.europa.eu/sites/edp/files/publication/18-01-25_eag_report_en.pdf; Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission (2019) Ethics Guidelines for Trustworthy AI. https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html.

6 See below Section C.

7 See European Parliament, P9_TA(2023)0236, Artificial Intelligence Act. Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))1, Article 4a.

based on delegation to standard-setting bodies that do not have the appropriate profile to deal with fundamental rights.

B. The AI Act: two reasons for a compromise solution

In the early stages of the Industrial Revolution, explosions and fires were common due to the limited ability to control steam power at the time. Similarly, the large-scale production of goods resulted in a number of defective products that harmed their users. In both these scenarios, the most effective response of the legal system in order to protect the injured parties would have been to introduce a strict liability regime to minimise the side effects of innovation and industrial development. However, it was only when these technologies reached a higher level of maturity and it became easier and cheaper to put in place safety measures to prevent their side effects that fault-based models were replaced by stricter forms of liability.

Decades later, at the dawn of the information society era, both US and EU legislators decided that it was better to limit the liability of Internet service providers, despite the fact that online BBSs and webpages hosted illegal or defamatory content. The rise of dominant platforms and their better position (and wider availability of resources) in content management later changed the initial scenario and recently led lawmakers to set specific obligations focused on competition, consumer protection and fundamental rights.[8]

Other examples of the relationship between technological development and regulation could be added. However, the pattern remains the same: the early stages of implementation of innovations require a kind of 'tolerance' from the legal system, accepting a certain degree of side effects on individuals and society in return for future benefits from investment in new technologies.

In addition, limiting the legal requirements for innovative technologies facilitates the entry of more players into the new industry and increases the investment of major players. Both of these effects contribute to a more

8 See Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) and Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act).

mature technology and, ultimately, to an easier reduction of side effects, which makes it possible to adopt stricter liability rules.

It is only by taking into account this logic, based on 'the gift of the evil devil',[9] that it is possible to understand the main reason for the regulatory approach adopted by the EU legislator in addressing the issues raised by AI. The minimalist approach to the protection of individual rights and society in the AI Act – as shown by its focus on only the most dangerous scenarios, i.e. prohibited and high-risk applications – clearly contradicts the earlier debate on the ethics of AI (see Section C), the Council of Europe's early work on the core role of human rights in AI development, and the growing interest in human rights in business regulation. However, it represents a compromise solution in dealing with the risks of a promising new industry sector.

Setting some redlines and introducing a general reference to the impact on human rights is rather far from the idea of a human-centred AI, and departs from the debate of recent years on the empowerment of citizens in the digital society. In the same way, a regulation based on the traditional industrial risk approach (conformity assessment, standards, market surveillance) is rather far from a principles-based and risk-focused law centred on human rights, such as the GDPR.

Looking at the GDPR, a long-celebrated EU success in digital regulation due to its global impact,[10] and comparing it to the AI Act, the different approaches become clear. While the GDPR has addressed potential personal data processing concerns by providing some principles along the line of a cautious approach to the use of personal information (e.g. purpose limitation, minimisation, storage limitation), the AI Act does not set any principles to guide AI development,[11] focusing only on risk mitigation. The emphasis on European values to be embedded in AI (although human rights are more properly universal values), which characterised the initial academic and regulatory debate, has been lost in favour of a risk-centred and mainly safety-oriented regulation.

9 See Calabresi, Guido. 1985. Ideals, Beliefs, Attitudes, and the Law: Private Law Perspectives on a Public Law Problem (Syracuse, New York: Syracuse University Press) 1985, Ch. 1.

10 See, e.g, Greenleaf, Graham, Now 157 Countries: Twelve Data Privacy Laws in 2021/22 (March 15, 2022). (2022) 176 Privacy Laws & Business International Report 1, 3-8, UNSW Law Research, Available at SSRN: https://ssrn.com/abstract=4137418.

11 But see the European Parliament's proposal in this regard below in Section D.IV below.

Nor is the risk-based approach comparable to the way in which risk has been framed over the years in data protection, i.e. the main regulatory framework of the digital society so far. While Article 35 of the GDPR takes a broad approach to risk identification and is rather demanding in risk mitigation, the AI Act is less strong in this direction. The paradigm shift is evident in the different way of addressing the key points of risk management, namely the classification of high-risk cases and the consequences of such classification.

Although both the regulations focus on high risk, the AI Act sets out a closed list of cases classified as high risk, while the GDPR leaves the obligation to assess the level of risk for each operation to the duty bearers (i.e. data controllers). The difference is significant: a closed list is less effective in identifying high-risk cases in the context of a rapidly evolving technology such as AI at this stage, as confirmed by the debate on LLMs and the 'last minute' amendments to the AI Act.[12] Moreover, Annex III of the AI Act identifies high-risk areas using rather broad descriptions that may include cases that are not high-risk.[13] This is only partly mitigated by the proposed 'reasoned notification' to the competent national supervisory authority,[14] which will result in a cumbersome process and does not preclude future litigations on the applicability of AI to specific products/services.

Looking at the legislative process and its democratic legitimacy, the Commission's power to amend this list raises some concerns about the unilateral role the Commission will play, given the direct impact of listed cases on the scope and applicability of the AI Act.

12 See European Parliament, P9_TA(2023)0236 (fn. 7), Article 28b (Obligations of the provider of a foundation model). See also Bertuzzi, Luca. «Leading EU Lawmakers Propose Obligations for General Purpose AI». www.euractiv.com, 14 March 2023. https://www.euractiv.com/section/artificial-intelligence/news/leading-eu-lawmakers-propose-obligations-for-general-purpose-ai/.

13 For example, the category of AI systems "intended to be used for the purpose of assessing students in educational and vocational training institutions" (Annex III, AI Act, Commission Proposal) may also include AI-supported examination systems that automate some assessment procedures, but without involving high risk.

14 See European Parliament, P9_TA(2023)0236 (fn. 7), Article 6 (2a) ("[providers] shall submit a reasoned notification to the national supervisory authority that they are not subject to the requirements of Title III Chapter 2 of this Regulation. […] Without prejudice to Article 65, the national supervisory authority shall review and reply to the
notification, directly or via the AI Office, within three months if they deem the AI system to be misclassified").

If we look at the consequences of this high-risk classification, they are necessarily milder in the AI Act than in the GDPR. Whereas in the later the non-negotiable protection of human rights led the EU legislator to put data processing applications entailing a high risk to individual rights off the market (Articles 35.7.d and 36.1), in the AI Act the legislator opted for an 'acceptable' risk, which means that risky applications can be used even though the level of risk remains high.

Finally, the AI Act does not provide clear criteria for a methodology to assess the impact on fundamental rights, as discussed further in Section E. In this respect, the similarity with the GDPR is only apparent. While it is true that Article 35 of the GDPR does not set out a methodology for DPIA, it is worth noting that the GDPR builds on more than four decades of data protection regulation and practice, during which several robust methodologies have been developed, starting with PIA models.[15] On the contrary, the AI Act builds on Human Rights Impact Assessment, which has only recently been developed in the business sector and does not fit properly with AI applications.[16] Nor does the idea of delegating the definition of this methodology to standards and standardisation bodies seem any more promising (see Section E).

This brief comparison between the AI Act and the GDPR shows how different the maturity of these two regulations is, as well as the different maturity of the industry they regulate. While several generations of data protection laws preceded the GDPR,[17] gradually increasing the level of protection hand in parallel with the development of more privacy-enhancing technologies, the AI Act is a first regulation at an early stage in the development of the AI sector on a large scale.

Recalling Reidenberg's six ways of shaping technology,[18] the EU cannot use the 'bully pulpit', has limited resources to fund AI innovation, and –

15 See also Wright D, De Hert P (eds) (2012) Privacy Impact Assessment. Springer, Dordrecht.

16 While traditional Human Rights Impact Assessment (HRIA) models are usually territory-based, considering the impact of business activities in a given local area and community, in the case of AI applications this link with a territorial context may be less significant. For a broader analysis of HRIA in AI see Mantelero. Beyond Data (fn. 4), Ch. 2.

17 See Mayer-Schönberger V (1997) Generational Development of Data Protection in Europe. In: Agre PE, Rotenberg M (eds) Technology and Privacy: The New Landscape. The MIT Press, Cambridge, pp 219–241.

18 Joel R. Reidenberg, 'Lex Informatica: The Formulation of Information Policy Rules Through Technology', Texas Law Review 76, no. 3 (1998): 553–84.

due to the fragmentation of its national and regional strategies – it faces some difficulties in using participation and bargaining power in procurement to shape an AI industry dominated by non-EU players. Regulation is therefore the main way in which the EU can influence the design of AI products and services provided to EU citizens and users.

However, regulating a market with weak regional champions necessarily requires a more industry-friendly approach than in the case of a more balanced market composition. For this reason, the first generation of EU regulation on AI must combine safeguarding the development of the AI industry with a minimum level of consistency with the EU's fundamental rights framework.

As was the case with the Industrial Revolution and the Internet revolution, it is not surprising that in the AI revolution the first regulatory framework only partially addresses the demand for the protection of individual and societal rights. This is why the high-level commitment to ethics and human rights of the early AI debate in Europe has more pragmatically ended up in an industry-focused regulation, centered on conformity assessment, with limited emphasis on fundamental rights.

However, as has been the case in other fields and given the rapid development of the AI industry, it is worth considering alternative paths that have now been discarded, but which may be part of the further development of the AI regulation or complement its implementation. From this perspective, the following section focuses on the role that ethics and human rights could play in a more holistic and mature AI regulation, which could represent the next horizon for an EU model with the ambition to replicate the so-called Brussels effect achieved with the GDPR.

C. The solutions left behind: an ethical and socially conscious approach, a principles-based model focused on human rights

Considering the alternative paths that the EU legislator has left open when framing the first AI regulation is not just a theoretical exercise, but a way to reflect on possible options to improve a regulation that does not fully address the main concerns about the impact of AI on society.

On the one hand, the focus on risk/conformity assessment reveals a techno-solutionist approach. Simply mitigating risks to make them acceptable is far less than creating a framework to support developers in shaping

human-centred AI that addresses the challenges AI poses to human rights, societal and ethical values.[19]

This is not only a general issue, but also an important part of the EU debate on AI regulation before the Commission set a different paradigm with the AI Act proposal. A similar path can be seen in the work of the Council of Europe, where the initial broad approach focused on ethical issues and human rights has been replaced by a more risk-focused approach.[20]

The outcome of the EDPS expert group,[21] the guidelines of the Independent High-Level Expert Group on Artificial Intelligence,[22] and the first draft of the EU Parliament on AI regulation, which refers to ethical values,[23] have clearly taken a different and more holistic view of regulating the impact of AI on society. Although current industrial policy issues have led to a different outcome in the AI Act, limiting the regulation of a technology so relevant to societal change to risk management, this does not seem entirely in line with EU values and orientation (see also Section D).

In this respect, although the AI Act does not take into account the ethical and social consequences of the use of AI, the reflections and proposals elaborated in recent years should be considered in order to complement this industry-focused regulation. This can be done by translating the reflections on the societal aspects of AI into best practices that can provide specific tools for value-oriented AI design and thus fill the gap in this first generation of AI regulation.

In 2015, facing the challenges of Big Data, IoT and cloud computing (three core components of the latest AI revolution), the EDPS considered

19 On the social and ethical component of AI systems design see also Mantelero. Beyond Data (fn. 4), Ch. 3, for further discussion and references.

20 Alessandro Mantelero and Francesca Fanucci. 2022. The International Debate on AI Regulation and Human Rights in the Prism of the Council of Europe's CAHAI: Great Ambitions. In: European Yearbook on Human Rights 2022 / Czech P., Heschl L., Lukas K., Nowak M., Oberleitner G., Cambridge, Intersentia, pp. 225-252.

21 See European Data Protection Supervisor, Ethics Advisory Group (2018) Towards a digital ethics. https://edps.europa.eu/sites/edp/files/publication/18-01-25_eag_report_en.pdf.

22 Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission (2019) Ethics Guidelines for Trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

23 See, e.g., European Parliament. 2020. Draft Report with recommendations to the Commission on a framework of ethical aspects of artificial intelligence, robotics and related technologies (2020/2012(INL)). See also Matos Pinto, Inês de. «The Draft AI Act: A Success Story of Strengthening Parliament's Right of Legislative Initiative?. 22(4) ERA Forum 2021: 619–41. https://doi.org/10.1007/s12027-021-00691-5.

that "an ethical framework needs to underpin the building blocks of this digital ecosystem"[24] and set up an Ethics Advisory Group (EAG) to open the debate on the ethical dimension of data-intensive technologies. This group of experts emphasised that the challenges posed by these technologies had only been partially addressed by the law and "ethics allows this return to the spirit of the law and offers other insights for conducting an analysis of digital society, such as its collective ethos, its claims to social justice, democracy and personal freedom."[25] Rejecting an instrumental approach to ethics, based on ethical checklists and a set of measures, the EAG encouraged "proactive reflection about the future of human values, rights and liberties, including the right to data protection, in an environment where technological innovation will always challenge fundamental concepts and adaptive capabilities of the law".

Despite a critical overlap between ethical and legal values, the outcome of the EAG clearly highlighted the tension between the challenges posed by data-intensive technologies and the response provided by the law, where the latter only partially addresses the diversity of societal consequences.[26] Responsible innovation and value-sensitive design, based on co-shaping of ethical considerations and design solutions in a case-by-case approach, were proposed[27] as methodological path towards digital ethics.

Unfortunately, this focus on methodology was largely neglected in the next widely promoted initiative of the European Commission, the Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission.[28] Leaving aside certain critical issues in the compo-

24 European Data Protection Supervisor. «Opinion 4/2015 Towards a new digital ethics», 2015, 12. https://edps.europa.eu/sites/default/files/publication/15-09-11_data_ethics_en.pdf.

25 European Data Protection Supervisor, Ethics Advisory Group (2018) Towards a Digital Ethics, 7.
https://edps.europa.eu/sites/edp/files/publication/18-01-25_eag_report_en.pdf.

26 European Data Protection Supervisor, Ethics Advisory Group (fn. 11), 15 ("The new digital age generates new ethical
questions about what it means to be human in relation to data, about human knowledge and about the nature of human experience. It obliges us to re-examine how we live and work and how we socialise and participate in communities. It touches our relations with others and perhaps most importantly, with ourselves").

27 European Data Protection Supervisor, Ethics Advisory Group (fn. 11), 22.

28 Independent High-Level Expert Group on Artificial Intelligence set up by the European
Commission (2019) Ethics Guidelines for Trustworthy AI. https://ec.europa.eu/futurium/en/aialliance-consultation.1.html.

sition and working methodology of this group[29] and (as with the EAG) the overlap between ethical and legal values, the main output of this group (The Assessment List For Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment[30]) is a questionnaire-based self-assessment tool.[31]

Although ALTAI emphasises the importance of a multidisciplinary team in carrying out the assessment,[32] the proposed model only provides only a few questions on societal impact with a very narrow focus,[33] unable to address the wide range of societal consequences of the use of AI in many fields.

This limitation and the long list of questions, which only partially address ethical and societal issues,[34] show the inherent weaknesses of using a questionnaire-based model to address these issues, whereas relying on

29 See also Thomas Metzinger. 2019. Ethics washing made in Europe, available at https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html, accessed on April 11, 2019. Thomas Metzinger was a member of the expert group.

30 Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission. «The Assessment List For Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment», 2020. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

31 The overlap with legal values is evident in the questions on fundamental rights, privacy and data governance, technical robustness and safety, diversity, non-discrimination and fairness, and environmental impact. The self-assessment checklist also includes questions on AI and risk management, such as those on technical robustness and safety, transparency (traceability), and accountability. The latter refers in part to legal issues, where accountability questions relate to auditing and redress in the event of harm.

32 Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission. «The Assessment List For Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment» (fn. 16), 4.

33 These are the proposed questions: Could the AI system have a negative impact on society at large or democracy?; Did you assess the societal impact of the AI system's use beyond the (end-)user and subject, such as potentially indirectly affected stakeholders or society at large?; Did you take action to minimize potential societal harm of the AI system?; Did you take measures that ensure that the AI system does not negatively impact democracy?

34 Some ethical and social issues are considered by the questions of the assessment model in the following areas: human agency and autonomy, human oversight (this part also includes questions on risk management), transparency (with regard to the question on traceability of the outcomes of the algorithmic system and some questions on explainability and communication, which are ancillary to stakeholder and right holder participation), diversity, non-discrimination and fairness (as far as universal design and shareholder participation are concerned), and accountability (limited to the question on the establishment of an ethics review board).

expert evaluation combined with a participatory approach is more appropriate, as initially pointed out by the EDPS.[35] In addition, reducing ethics to a checklist largely limits the consideration of ethical issues by turning them into a functional risk-centred analysis with a very limited role for value design.

Based on this experience, some suggestions can be made for the future implementation of the AI Act or, more generally, for a future holistic approach to AI.

First, a regulatory model centred on conformity assessment, while including human rights, leaves out important issues that need to be taken into account in the implementation of AI projects. AI applications cannot be considered as given, unquestionable and only assessed to minimise their high risk. First of all, it is necessary to examine their social acceptability and, if confirmed, the way in which societal values (including ethical values) are embedded in the solutions adopted while avoiding conflicts with the values of the target communities. Real cases have shown that projects are likely to fail if these aspects are ignored.[36]

The AI Act should therefore be accompanied by appropriate solutions to integrate ethical and societal issues into the evaluation of potential uses of AI. The AI Act should not be seen as the end of the ethical debate: while it makes a positive contribution to avoiding improper overlaps between ethics and law, it does not address the societal issues that need to be faced before the development of an AI capable of passing the AI Act's conformity assessment. In this respect, acceptability is not just a matter of safety, security and respect for fundamental rights.

35 On participation see also European Center for Not-for-Profit Law Stichting (ECNL). 2023. «Framework for Meaningful Engagement: Human Rights Impact Assessments of AI | ECNL», https://ecnl.org/publications/framework-meaningful-engagement-human-rights-impact-assessments-ai; Data Justice Lab. 2022. «Civic Participation in the Datafied Society: Towards Democratic Auditing?», https://datajusticelab.org/wp-content/uploads/2022/08/CivicParticipation_DataJusticeLab_Report2022.pdf; Mantelero. Beyond Data (fn. 4), 127-130.

36 See, e.g., Scassa T (2020) Designing Data Governance for Data Sharing: Lessons from Sidewalk Toronto.
Technology & Regulation, Special Issue: Governing Data as a Resource, Technology and
Regulation 44–56; Goodman EP, Powles J (2019) Urbanism Under Google: Lessons from Sidewalk Toronto.
Fordham L. Rev. 88(2):457–498.

Second, in putting the ethical and societal impact of AI at the heart of the debate, it is important to avoid turning the assessment exercise into a mere technical tool, where an in-depth analysis of guiding values and their integration into AI solutions is replaced by standardised questions. As in the long-lasting experience of ethics boards and committees, including some recent implementations of this practice in the AI industry, it is important to understand individual and societal needs and be able to mediate them through technology, building a value-oriented AI design that is aligned with the characteristics of the context in which AI will be deployed.

Three elements are crucial to achieve this result: independent experts, the commitment of AI developers, and the active engagement of the community where AI solutions will be implemented. The latter are not necessarily territory-based communities but often large and distributed communities of end users.

Without repeating considerations expressed elsewhere, it is worth noting that societal impact assessment is more complicated than human rights impact assessment because it cannot benefit from a well-defined and, to a large extent, universal benchmark.[37] The key elements are therefore contextualisation, based on expert insights into the values to be taken into account, and participation, which is useful to complement and verify this expert assessment.

There is no one-size-fits-all way to implement this model centred on these two elements, and we should also be aware that the structure, composition and internal organisation of expert committees are not neutral elements, nor is the way in which stakeholders and rightsholders are involved. In both cases, the manner in which the assessment is carried out influence its outcome in terms of quality and reliability of the results.

Given the contextual nature of the AI projects and their impacts, it is possible to identify some key elements that characterise these expert bodies (e.g. independence, multidisciplinarity, and inclusiveness; transparency of internal procedures and decision-making processes; provisional nature of their decisions), but a variety of structures and types of organisation are possible in terms of (i) member qualifications, (ii) rights-holder, stakeholder, and layperson participation, and (iii) internal or external experts.

With regard to the commitment of AI developers, it is important to build a bridge between these expert bodies and the day-to-day activities of

37 See fn. 19.

AI development. In this respect, an ethical body or advisory team should neither be seen as a control body, making it difficult to accept its role, nor as a body to which all ethical and social issues can be delegated, with the implicit lack of a by-design approach by developers from the early stages of AI projects.[38]

Finally, regarding the role of the participatory approach in dealing with societal issues, it is worth emphasising that participation not only contributes to a better understanding of the societal and ethical issues, but is also essential for effective democratic decision-making in AI.

D. In search of a principles-based core for AI regulation

Looking at the framework for the development of AI set out by the EU legislator in the AI Act, it is not only the societal issues that are critical, but also the way in which the focus on fundamental rights has been framed.

In other crucial areas of technological development, such as biotechnologies and digital information, the European legislators have usually developed more elaborate regulatory instruments that establish a set of principles to guide operators in shaping technology, rather than simply affirming human rights and societal values. This was the case with the Oviedo Convention and the EU regulation and practice on clinical trials, as well as in the case of the information society where Convention 108+[39] and the GDPR set out guiding principles to embed key values in the design of medical, biomedical, and ICT products and services.

A general part focusing on key principles was absent from the debate on the AI Act and was only proposed at the end in the amendments adopted by the European Parliament (see Section D.IV). However, given the pervasive nature of AI and the wide range of its applications, defining a set of common principles is not an easy task.

In the following sub-sections three different contexts are considered in the search for possible guiding principles. First, the main principles that

38 One possible solution is to appoint an internal advisor on societal issues (also known as Chief Ethics Officer) as a permanent contact for day-to-day project development and as a trait d'union with the external experts.

39 See also Mantelero, Alessandro; Stalla-Bourdillon, Sophie, and Kwasny, Sophie (eds). 2021. Convention 108 and the future data protection global standard. Computer Law & Security Rev., Special Issue, https://www.sciencedirect.com/journal/computer-law-and-security-review/special-issue/10FW5NWHJFK.

have emerged in the ethical debate on AI will be considered, as they often have a legal dimension. This is followed by an examination of two other specific initiatives: the Council of Europe's draft framework convention on AI and the principles set out by the NIST and the Blueprint for an AI Bill of Rights in the US. The potential impact of these two initiatives on global trends in the regulation of AI and in the definition of its core principles is related to the international scope of the Council of Europe's approach and the prominent position of US companies in AI development, respectively.

I. The principles set out in the ethical charters

Growing concerns about the impact of AI on individuals and society have stimulated a wide range of initiatives to outline key guiding values for AI development.[40] Looking at this corpus of ethical charters can provide some suggestions for relevant values to be included in AI regulation, to be translated into legal values or be considered as part of the legal assessment, as is the case in the regulation of biomedicine and scientific research.

Several studies[41] have focused on the key values of these guidelines identifying a small core of values that are common to most of the documents. according to a first study,[42] five of them are ethical values with a strong

40 See also Raab, Charles D. «Information Privacy, Impact Assessment, and the Place of Ethics». Computer Law & Security Review 37 (6 March 2020): 105404. https://doi.org/10.1016/j.clsr.2020.105404 ("a bewildering array of ethics boards, panels, committees, groups, centres, frameworks, principles, templates, guidelines, protocols, projects and the like have all popped up like woodland mushrooms in a wet Autumn").

41 It is worth pointing out some of the limitations of these studies: the use of grey literature, the use of search engines for content selection, linguistic biases, and a quantitative text-based approach that underestimates the policy perspective and contextual analysis. From a policy and regulatory perspective, their main limitation is the quantitative approach adopted, which considers differing sources at the same level, without taking into account the differences between the guidelines adopted by governmental bodies, independent authorities, private or public ad hoc committees, big companies, NGOs, academia, intergovernmental bodies etc. When the focus is on values for future regulation, the different relevance of the sources in terms of political impact is important, and the mere frequency of occurrence does not take this impact into account.

42 Jobin et al. 2019. The authors identified ten key ethical values within a set of 84 policy documents with the following distribution: transparency 73/84; non-maleficence 60/84; responsibility 60/84; privacy 47/84; beneficence 41/84; freedom and autonomy 34/84; trust 28/84; sustainability 14/84; dignity 13/84, and solidarity 6/84.

legal implementation (transparency, responsibility, privacy, freedom and autonomy) and only two come from the ethical discourse (non-maleficence and beneficence). Another study[43] identified several guiding values and the top nine are: privacy protection; fairness, non-discrimination and justice; accountability; transparency and openness; safety and cybersecurity; common good, sustainability and well-being; human oversight, control and auditing; solidarity, inclusion and social cohesion; explainability and interpretability. As in the previous study, the aggregation of these principles is necessarily influenced by the categories used by the authors to reduce the diversity of principles.

If we take a qualitative approach, limiting the analysis to the documents adopted by the main European organisations and those with a general and non-sectoral perspective,[44] we can better identify the key values that are most popular among rule makers.

looking at the four core principles[45] identified by the High-Level Expert Group on Artificial Intelligence,[46] respect for human autonomy and fairness are widely developed legal principles in the field of human rights and law in general, whereas explicability is a technical requirement rather than a principle. With regard to the seven requirements[47] identified by the HLGAI on the basis of these principles, human agency and oversight are further specified as respect for fundamental rights, informed autonomous decisions, the right not to be subject to purely automated decisions, and the adoption of oversight mechanisms. These are all requirements that are already present in the law in various forms, particularly in relation to data processing. The same applies to the remaining requirements (technical robustness and safety, privacy and data governance; transparency; diversity, non-discrimination and fairness; accountability; and environmental well-being).

43 Hagendorff (fn. 4), p 102.

44 E.g. Council of Europe – European Commission for the Efficiency of Justice (CEPEJ) 2018.

45 Respect for human autonomy, Prevention of harm, Fairness, Explicability.

46 Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission 2019.

47 Human agency and oversight; Technical robustness and safety; Privacy and data governance; Transparency; Diversity, non-discrimination, and fairness; Societal and environmental wellbeing; Accountability.

Another important EU document identifies the following nine core principles and democratic prerequisites:[48] human dignity; autonomy; responsibility; justice, equity, and solidarity; democracy; rule of law and accountability; security, safety, bodily and mental integrity; data protection and privacy; sustainability.

Based on the results of these different (quantitative, qualitative) methods of analysis, we can identify three main sets of values that are relevant from a regulatory perspective and can form a set of principles-based references for a more holistic implementation of AI regulation.

The first consists of the contextual application of principles that are already enshrined in law but play a crucial role in AI, such as privacy and data protection, fairness, non-discrimination, justice, freedom, and autonomy. A second group covers general legal principles that are relevant in the AI context and includes transparency, explainability, interpretability, accountability, and responsibility. The last group, which includes safety and cybersecurity, control and auditing, transparency, openness and human oversight, consists of principles that deal with technical and procedural issues.[49]

II. Principles identified by the Council of Europe

In 2019, the Council of Europe started a reflection on the adoption of a future convention on AI. On the basis of preliminary studies on the legal[50] and ethical dimensions[51] of AI regulation, the Ad Hoc Committee on Artifi-

48 European Commission - European Group on Ethics in Science and New Technologies 2018.

49 In a way that is consistent with the nature of these ethical charters, they also include specific ethical values derived from ethical and sociological theory (e.g., common good, well-being, solidarity) and principles from applied ethics and research/medical ethics (e.g., non-maleficence, beneficence). These principles can play a crucial role in addressing societal issues related to the use of AI, but need to be properly contextualised to avoid the potential risk of 'transplanting' of ethical values.

50 Mantelero A (2020) Analysis of international legally binding instruments. In Council of Europe. Towards regulation of AI systems. Global perspectives on the development of a legal framework on Artificial Intelligence systems based on the Council of Europe's standards on human rights, democracy and the rule of law. DGI (2020)16, pp 61–119.

51 Ienca M, Vayena E (2020) AI Ethics Guidelines: European and Global Perspectives. Ibidem, pp 38–60.

cial Intelligence (CAHAI) elaborated its feasibility study[52] and developed a participatory process among its members and with the involvement of external stakeholders. This led to a first draft of the Possible elements of a legal framework on artificial intelligence, based on the Council of Europe's standards on human rights, democracy and the rule of law.[53]

After this first phase of the drafting process, a new committee (the Committee on Artificial Intelligence - CAI) took over from the CAHAI with the task of providing an "appropriate legal instrument on the development, design, and application of artificial intelligence systems based on the Council of Europe's standards on human rights, democracy and the rule of law, and conducive to innovation, in accordance with the relevant decisions of the Committee of Ministers".[54]

This analysis focuses on the Revised Zero Draft [Framework] Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law, which was prepared by the Chair of the CAI with the assistance of the Secretariat to serve as a basis for the drafting of the future convention on AI and "does not reflect the final outcome of negotiations in the Committee".[55]

The proposed draft included six guiding principles: equality and non-discrimination; privacy and personal data protection; accountability and

52 Council of Europe, Ad hoc Committee on Artificial Intelligence (CAHAI). 2020a. Feasibility
Study, CAHAI(2020)23. https://rm.coe.int/cahai-2020-23-final-eng-feasibility-study-/1680a0c6da.

53 The text of the proposal is available here https://rm.coe.int/possible-elements-of-a-legal-framework-on-artificial-intelligence/1680a5ae6b. For a critical analysis of the CAHAI work and results, see also Mantelero, A. and Fanucci, F. 2022. Great ambitions. The international debate on AI regulation and the human rights in the prism of the Council of Europe's CAHAI. In Philip Czech et al. (eds). European Yearbook on Human Rights 2022 (Intersentia: Cambridge), pp. 225-252.

54 See the CAI's Terms of Reference available here: https://rm.coe.int/terms-of-reference-of-the-committee-on-artificial-intelligence-for-202/1680a74d2f.

55 Adopted in Strasbourg, on 6 January 2023 CAI(2023)01, and available here https://rm.coe.int/cai-2023-01-revised-zero-draft-framework-convention-public/1680aa193f.
At its 4th Plenary meeting, the CAI decided to make the revised "Zero Draft" [Framework] Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law public. A more recent document was prepared by the Chair of the CAI, see Committee on Artificial Intelligence. Consolidated working draft of the Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law, Strasbourg, 7 July 2023 CAI(2023)18, https://rm.coe.int/cai-2023-18-consolidated-working-draft-framework-convention/1680abde66 (last accessed 11.08.23). This document does not reflect the outcome of the negotiations in the Committee and is therefore not considered here as the main reference.

responsibility; transparency and oversight; safety; safe innovation. Apart from the last one (safe innovation), which is not a principle but only a provision to legitimise the use of the so-called regulatory sandbox for AI, the others mainly recall existing principles[56] without any specific contextualisation with regard to AI.

In addition, accountability, responsibility and legal liability, as well as safety, cannot be considered as principles but as general rules or operational legal requirements. Incidentally, it is worth noting that their specific application in the field of AI is much debated, both in terms of how to allocate AI liability and how to ensure safety through standards or other solutions.[57] Against this background, a general reference to these criteria does not provide any specific regulatory guidance to the states.

Only the specific requirement to develop "adequate oversight mechanisms as well as transparency and auditability requirements tailored to the specific risks arising from the context in which the artificial intelligence systems are applied are in place" seems to provide a specific contribution to AI regulation in terms of broad transparency and oversight obligations.

Building on the Council of Europe's legal framework, a different approach could have been adopted by contextualising the principles already enshrined in the legal instruments of the Council of Europe in relation to the challenges posed by AI. For example, the principle of beneficence enshrined in Article 6 of the Oviedo Convention[58] can be applied to AI in a context-specific way, where the complexity or opacity of AI-based solutions places limits on individual consent, which therefore cannot be the exclusive basis for intervention.[59]

Comparing the Revised Zero Draft with the Oviedo Convention and Convention 108/108+, the difference between conventions that establish a

56 See, e.g., Article 12 of the Revied Zero Draft which merely states that "Each Party shall, within its jurisdiction and in accordance with its domestic law, ensure that the design, development and application of artificial intelligence systems respect the principle of equality, including gender equality and rights related to discriminated groups and individuals in vulnerable situations".

57 See European Commission, Proposal for a Directive of the European Parliament and the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive). COM/2022/496 final. See also The European AI Liability Directives – Critique of a Half-Hearted Approach and Lessons for the Future, Computer Law and Security Review, forthcoming, https://arxiv.org/abs/2211.13960.

58 See also Oviedo Convention, Articles 16 and 17.

59 For a proposal in this regard, see Mantelero Beyond (fn. 4), Ch. 4, para 4.2, and Mantelero. Analysis of international legally binding instruments (fn. 50).

specific framework of principles focused on their specific scope and the CAI proposal, which simply recalls for the respect for existing rights and principles without any contextualisation to address the challenges of the AI environment, is clear.

III. The principles set out by the US National Institute of Standards and Technology (NIST)

The NIST framework is characterised by a peculiar approach to risk, as the report clearly states (emphasis in the original text): "While risk management processes generally address negative impacts, this Framework offers approaches to minimize anticipated negative impacts of AI systems *and* identify opportunities to maximize positive impacts". In defining these opportunities, the Framework refers to "potential benefits to people (individuals, communities, and society), organizations, and systems/ecosystems".[60]

These general statements about the core of the risk assessment model are consistent from a risk analysis perspective, but raise some key issues from a legal perspective, which become relevant when – as in the AI Act – risk assessment is part of a regulatory compliance framework with associated obligations, sanctions for non-compliance, and potential liability.[61]

The first main issue concerns the decision to include benefits in the risk assessment. Leaving aside the difference between a purely risk-based approach and a rights-based approach to risk, a key issue is who is entitled to define the "benefits to people (individuals, communities, and society), organizations, and systems/ecosystems" that may justify exposure to risk, including potential prejudice to human rights.

The balancing of competing interests is common in law, but is based on a legal assessment that, in accordance with the relevant legal system, wights the interests of individuals, communities, and society as defined through a democratic process that results in legal provisions and their interpretation by the courts.

Here, in the Framework, this balancing exercise between the negative impacts of the use of AI – which includes restrictions or prejudice to individual and collective rights and freedoms – and its benefits is carried out

60 National Institute of Standards and Technology. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0), https://doi.org/10.6028/NIST.AI.100-1, 4.

61 See also European Commission, AI Liability Directive (fn. 57) and its relation to the AI Act.

by AI developers outside any democratic and participatory framework. In short, a company will decide what are the individual/collective benefits and restrict individual/collective rights and freedoms without any legitimacy.

Different considerations could be made for the public sector, where the nature of public bodies and their mandate may legitimise them to conduct an assessment of individual and collective interests based on the power vested in them by law and the associated democratic scrutiny of the exercise of that power.

Against this background, while assessing the negative impacts of AI is an exercise that can be carried out by AI developers as the framework is given (i.e. human rights, mandatory security and conformity rules and principles), this framework is not given for the potential benefits. Given the trade-off between benefits and negative impacts, and the variety of potential benefits – which could include purely economic benefits – this exercise differs from the traditional balancing test between competing interests protected by law, and opens the doors to self-assessment by AI developers, who end up deciding what the societal benefits of AI are.[62]

Although this concern can be mitigated by the participatory approach proposed by the Framework,[63] which involves experts and civil society in the evaluation, the lack of a model for participatory assessment[64] seriously hampers the possibility of using this broader engagement to mitigate the concerns outlined above.

The NIST document sets out some principles for AI development, but – with some exceptions on fairness and privacy – they do not focus on societal needs in relation to AI and the legal and societal values that should underpin AI development and its use. They are mainly technical requirements, according to which an AI system must be valid, reliable,[65]

62 These critical considerations can also be extended to the amendment proposed by the European Parliament regarding the risk management system in the AI Act, see European Parliament, P9_TA(2023)0236, Article 9.5 ("High-risk AI systems shall be tested for the purposes of identifying the most appropriate and targeted risk management measures and weighing any such measures against the potential benefits and intended goals of the system. Testing shall ensure that high-risk AI systems perform consistently for their intended purpose and they are in compliance with the requirements set out in this Chapter").

63 See pp. 11, 17, and 24.

64 See p. 24.

65 i.e. ability of an item to perform as required, without failure, for a given time interval, under given conditions.

robust,[66] safe, secure, resilient, accountable, transparent, explainable and interpretable.

Meeting these technical requirements can certainly help to design of a human-centered AI that is more respectful of individual and collective rights, as well as the values and needs of society, but they are not in themselves capable to pave the way for value-oriented design that outlines the goals and boundaries of AI use in our society.

IV. A late addition: the European Parliament's general principles applicable to all AI systems

In the last round of amendments to the AI Act, the European Parliament tried to fill the gap in the act regarding to the lack of guiding principles for AI development. Although this was a critical shortcoming, the solution of copying the principles outlined by the Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission[67] does not fully address the criticisms that characterise the AI Act on this issue.

First, these principles were the result of a controversial drafting process.[68] Second, as noted above, they largely repeat requirements that already exist in various forms in the law. Although the proposed provision introduces some contextualisation of these principles, a broader analysis based on the main existing international legal instruments[69] could have helped to outline a more comprehensive framework of principles.

Looking at the individual principles, some of them seem superfluous. This is the case with the principle of 'privacy and data governance', which simply refers to existing privacy and data protection rules. It is also the case with the principles of 'diversity, non-discrimination and fairness' which refers to the promotion of gender equality and cultural diversity, and the prevention of discriminatory effects and unfair biases prohibited by Union

66 i.e. ability of a system to maintain its level of performance under a variety of circumstances.

67 See above Section D.I. There are still some misunderstandings and improper overlap between legal and ethical perspectives in the amendments proposed by the European Parliament, see e.g. European Parliament, P9_TA(2023)0236 (fn. 8), Recital 9a in relation to Article 4a.

68 See Thomas Metzinger. «Ethics washing made in Europe». Tagesspiegel 8 April 2019, https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html.

69 See fn. 50.

or national law. Here, the only element specifically relevant as a guiding principle for AI development and use is inclusiveness ("AI systems shall be developed and used in a way that includes diverse actors").

Similar considerations can be made with regard to the principle of 'human agency and oversight' where it requires AI systems to be developed and used "as a tool that serves people, respects human dignity and personal autonomy", which are either vague references (serving people) or general principles already enshrined in the EU framework.

Other principles are vague and cannot easily be adapted to the legal framework or contextual AI applications. This is the case with the principle of social and environmental well-being. On the one hand, the reference to sustainability and an environmentally friendly approach seems redundant within the broader EU legal framework, also in view of the rough description given for this principle. On the other hand, the commitment "to benefit all human beings, while monitoring and assessing the long-term impacts on the individual, society and democracy" sounds too ambitious and inconsistent with the uses of AI which, by their very nature, do not necessarily benefit all human beings and may also produce outcomes that adversely affect certain individuals, including in terms of legal effects.

To a large extent, this list of principles is mainly a vademecum, recalling principles and values that are already present in EU law or, if new, that can hardly guide AI developers and deployers due to their vagueness and wide scope.

The principles of transparency, technical robustness and safety are an effective contribution in terms of contextualisation of already existing general principles, especially with regard to the requirement for an AI design that allows for appropriate human control and oversight.

Even in this case, while technical robustness and safety are appropriately and contextually framed, the transparency principle refers to traceability and explainability but both these two requirements can be implemented in many different ways, leaving a wide margin for manoeuver to operators.[70]

While the definition of general principles is always a difficult exercise, in balancing the need for sufficient detail on their content with the nature of general principles, the list proposed by the European Parliament seems to be of limited help in guiding those who have to design, develop and deploy

70 Only a few obligations are listed specifically and in a rather general way (awareness of human-AI interaction, notification of the capabilities/limitations of the individual AI system, information on user rights).

AI products and services. Principles on aspects such as individual and community participation in AI design, public debate on the need to choose an AI-based solution over other possible options,[71] respect for community values and diversity in AI applications with a social impact are just some of the possible improvements for a human-centric approach to AI.

E. Towards a full implementation of the risk-based approach: the role of fundamental rights

Over the years, various technologies with a major impact on society and innovation have been regulated through international and regional instruments, establishing common principles and rights to pave the way for innovation in a manner consistent with societal values and aspirations.

These legal instruments have not only established general principles, nor have they simply emphasised the importance of ensuring human rights protection, but have also outlined guiding principles centered on the specific technological context to be regulated in order to support its values-oriented development, rather than being limited to its technical efficiency and safety nature.

Based on the brief analysis carried out in the previous sections, the ethical charters on AI, the Council of Europe's approach and the framework provided by the NIST have identified some common values for AI development (e.g. respect for equality and non-discrimination, protection of personal data, transparency, accountability, security), as has the European Parliament, but these are, to a large extent, general statements that are not contextualised in the specificity of the AI environment. It is therefore difficult to see in these various initiatives a clear approach capable of establishing principles that will effectively guide the development and use of AI.

Although a more contextual exercise was possible, it seems that at this early stage of AI regulation the elaboration of a tailored set of principles is not yet mature. In this regard, initiatives such as the European Declaration

71 See Council of Europe, Consultative Committee of the Convention for the Protection of Individuals
with regard to Automatic Processing of Personal Data (Convention 108) (2019) Guidelines on
Artificial Intelligence and Data Protection, T-PD(2019)01. https://rm.coe.int/guidelines-onartificial-intelligence-and-data-protection/168091f9d8, paras II.9 and III.8.

on Digital Rights and Principles for the Digital Decade may suggest that a longer journey towards a principles-based regulation is possible, but for now a pragmatic, risk-focused approach prevails.

This does not mean that, in the absence of guiding principles, legal issues related to AI only concern traditional industrial and product safety. On the contrary, the lack of adequate guiding principles call for the human centric-AI proposed by European legislators to be built on the existing human/fundamental rights framework. A framework that, even if not used to outline guiding principles, will necessarily feed into impact assessment models.

However, it is important not to overestimate the scope of the risk-based approach. An example of this is the Council of Europe's decision to develop an impact assessment model that can cover not only human rights (HRIA), but also democracy, and the rule of law. While the HRIA is not new and, with some modifications, can be used in AI, the inclusion of democracy and the rule of law is challenging in its transition from theoretical vision to concrete implementation.[72]

The democratic process and democracy in its various manifestations cover a wide range of issues, making it methodologically difficult to assess the impact of a technology and its various applications on them. Even more so since it is difficult to assess the level of democracy itself. This does not mean that it is impossible to carry out an impact assessment on specific areas of democratic life, such as the right to participation or access to pluralist information, but this remains a HRIA, albeit one centered on civil and political rights.

Different considerations apply to the rule of law, where the more structured field of justice and the limited application of AI make it easier to envisage uses and foresee their impact on a more uniform and regulated set of principles and procedures than in democracy. However, even in this case, the specificity of the field and the interests at stake may raise some doubts about the need for an integrated risk assessment model – encompassing human rights, democracy, and the rule of law – as opposed to a more limited assessment of the impact of specific AI applications on the rule of law.

72 The same consideration can be applied to the amended text of Article 9 adopted by the European Parliament, which has extended risk management to democracy and the rule of law.

While Human Rights Impact Assessment can be implemented in AI applications, it is still largely more a goal rather than a specific response provided by the legislators, who refer to it without providing concrete methodological solutions. As noted elsewhere, the traditional HRIA is primarily a policy tool rather than a regulatory tool, and it is usually territory-based and covers a wide range of rights and freedoms.[73] HRIA therefore needs to be reshaped to serve the purposes of AI regulation, which focuses on risk thresholds and risk management obligations.

In this respect, although the discussion and proposals on AI regulation are quite mature, the core issues relating to the model for carrying out such an assessment have not yet been worked out in such a way that provides meaningful input to companies and other actors that will have to comply with the AI Act.

Several limitations affect the proposed models: (1) use of lengthy questionnaires following an awareness-raising model rather than an impact assessment model;[74] (2) misunderstandings about the key parameters to be considered for impact assessment;[75] (3) an aggregate impact on fundamental rights regardless of their specific nature; (4) little focus on how to quantify the risk, which is at the heart of AI Act conformity assessment.

Based on the DPIA (Data Protection Impact Assessment)/PIA (Privacy Impact Assessment) experience, more streamlined proposals are needed providing a methodology rather than a fixed scheme that cannot cover the full range of AI applications. Self-assessment is possible but must be based on an independent expert evaluation, which is the only way to contextualise how specific AI applications may affect fundamental rights. An expert-based evaluation, combined with appropriate tools for stakeholder

73 See Mantelero. Beyond Data (fn. 4), Ch. 2.

74 See, e.g., Government of the Netherlands. 2022. Fundamental Rights and Algorithms Impact Assessment (FRAIA) https://www.government.nl/documents/reports/2021/07/31/impact-assessment-fundamental-rights-and-algorithms.

75 See, e.g., Government of the Netherlands (fn. 74), 74-75; The Alan Turing Institute. 2021. Human Rights, Democracy,
and the Rule of Law. Assurance Framework for AI Systems: A proposal prepared for the Council of Europe's Ad hoc
Committee on Artificial Intelligence, https://rm.coe.int/huderaf-coe-final-1-2752-6741-5300-v-1/1680a3f688 62-(where, e.g., the indices are calculated summing heterogeneous variables, using different scales, and including variables for the affected right-holders considered in total numbers rather than in proportion to the total number of the right-holders).

and rightsholder participation, can easily rely on a lean assessment model, which also improves the transparency of the assessment.

Such an approach is possible, not by using a 300-page impact assessment model, but by outlining the key elements to be considered and developing a methodology to combine and assess the different risk components of AI in relation to fundamental rights.[76]

Unfortunately, the current debate seems likely to complicate the way in which AI regulation, and the AI Act in particular, will be implemented. Unrealistic standards for HRIA/FRIA (Fundamental Rights Impact Assessment), rather than a general methodology, and lengthy and cumbersome assessment models (including misunderstandings in risk assessment) are not in line with the previous experience of DPIA/HRIA and are likely to turn FRIA into a box-ticking and bureaucratic exercise.

It would therefore be important for the EU bodies (including the European Union Agency for Fundamental Rights and CEN/CENELEC) to engage in a serious and inclusive discussion on FRIA, avoiding inner circles and including real domain experts and critical voices.

In this context, the European Parliament's proposal to introduce a Fundamental Rights Impact Assessment for high-risk AI systems[77] was intended to highlight the role of FRIA, but if not properly developed, FRIA risks to complicate the regulatory framework rather than addressing its limitations.

More specifically, the AI Act provides for three different forms of impact assessment: (i) a technological assessment based on a general evaluation of certain AI-based technologies in order to list them as high-risk applications (Annex III and Articles 6 and 7); (ii) a conformity assessment, focused on the specific AI applications and carried out by AI providers according to standards to be defined (Articles 9, 17 and 40); (iii) a FRIA, carried out by deployers, focused on the contextual use of the specific AI application and not based on standards. This combination of assessment procedures of different scope and nature gives rise to several internal conflicts.

First, the methodological criteria for carrying out the impact assessment are not defined or are (in the Parliament's text[78]) outlined in an unclear manner, focusing on the two main variables traditionally used in risk as-

76 These were the key elements of the methodology for HRIA in AI described in Mantelero. Beyond Data (fn. 4), Ch. 2, where a methodology for assessing the level of impact on the rights potentially affected by a given AI application is proposed.

77 See Article 29a of the text of the AI Act amended by the European Parliament.

78 See Article 3(1a) and 3(1b).

sessment (i.e., likelihood of harm and the severity of that harm) but adding other parameters in the definition of 'significant risk' – namely severity, intensity, probability of occurrence, duration of effects, ability to affect an individual, a plurality of persons or to affect a particular group of persons – which should be considered as sub-categories within the two main variables (e.g., the duration of negative effects should be included in severity).

Second, these three forms of assessment show a different approach to the assessment criteria and their setting: for the technology assessment to be carried out by the Commission (Annex III and Articles 6 and 7) the main variables are the severity and probability of occurrence of the risk;[79] no variables are provided for the conformity assessment to be carried out by AI providers, which is largely left to future harmonised standards (Article 40); no variables are provided for the performance of the FRIA by the natural or legal person, public authority, agency or other body using an AI system under its authority, and no standards are planned for this case (Article 29a).

Although, according to the general risk assessment theory, we can assume that the two criteria set for technology assessment (severity and probability) should be the same for the other forms of assessment, the lack of a clear guidance and the confusion in considering these criteria should lead the EU legislator to provide a common general framework of relevant parameters for impact assessment.

Moreover, there is a risk of methodological conflict in assessing the same AI application against specific standards (in the case of the AI provider) and pure self-assessment (in the case of the entity using an AI system under its authority), where the two methodologies may diverge. Nor does a generalisation of standards seem to be a better option.

The lack of experience of EU standardisation bodies in the field of fundamental rights,[80] their business-oriented composition and the lack of transparency in their procedures,[81] as well as the absence of an effective and

79 See Article 7 of the AI Act proposal.

80 See also European Commission, A Notification under Article 12 of Regulation (EU) No 1025/20121 on a standardisation request to the European Committee for Standardisation (CEN) and the European Committee for Electrotechnical Standardisation (CENELEC) in support of safe and trustworthy artificial intelligence, draft, Brussels, 5.12.2022.

81 Veale, Michael and Zuiderveen Borgesius, Frederik, Demystifying the Draft EU Artificial Intelligence Act (July 31, 2021). Computer Law Review International (2021) 22(4) 97-112.

broad democratic participation in standard-setting are clear limitations of this solution. Furthermore, it should be made clear that the standard for impact assessment should be a methodological standard that outlines the variables to be used in the assessment process and how they are quantified and combined: a methodological standard that is suitable for all the different contexts in which AI is used, and not just an awareness tool based on a long list of questions.[82]

Given the uniformity of risk assessment procedures based on the common theory of risk assessment, the easiest way to address the criticism discussed above is to provide a clear list of key parameters for risk assessment. These should be the same for all the forms of assessment, but with a different implementation in the technology assessment to be performed by the European Commission and in the contextual impact assessment to be carried out by providers and deployers, the former being a general ex ante evaluation based on case scenarios or similar tools, while the latter are contextual impact assessments related to a specific AI application. In addition, given the established elaboration of risk assessment and the previous experience in regulating data protection impact assessment, these criteria should be further implemented in a specific general methodology by AI supervisory bodies.

The feasibility and effectiveness of this approach is confirmed by the implementation of the GDPR, where supervisory authorities have played a key role in establishing uniform methodologies for DPIA. This solution, also in view of the composition of the proposed European Artificial Intelligence Office, can provide more competence, transparency and integration with the existing institutional bodies – including on fundamental rights issues – than delegation to standardisation bodies, which lack sufficient expertise and legitimacy to deal with fundamental rights issues.

The overall regulatory framework for risk assessment should therefore be based on three different layers: (i) common general criteria and variables to be used in impact assessment, defined for all types of assessment; (ii) impact assessment methodologies, defined by an ad hoc EU supervisory body (which also ensure EU-wide harmonisation); (iii) technical standards, set by standardisation bodies, covering the safety and security of the AI systems, but not their impact on fundamental rights.

As far as the AI providers and deployers are concerned, risk assessment should be an integrated tool with a proportionate distribution of burdens

82 See fn. 76.

based on the actual risk introduced into society and the ability of each actor to manage that risk, as generally accepted in the legal theory of risk. The AI provider should therefore carry out the initial assessment of a given AI product/service, taking into account all its potential uses, but the AI deployer using that product/service in a given context and for specific purposes should integrate this initial assessment with the analysis of the contextual impact and associated risks. This necessarily requires a flow of information between providers and deployers about the characteristics of the AI system and the risks associated with it.

Finally, a common methodology will facilitate not only the integration of assessments by AI providers and AI deployers, but also the integration of different AI products, thus making it possible to assess their cumulative risks.

F. Conclusions

Several initiatives around the world and at different levels are focusing on the regulation of AI. Lawmakers are trying to provide an first response to the challenges posed by the AI revolution. Focusing on the EU AI Act, this chapter has highlighted three main elements that characterise this first generation of AI laws.

First, the proposed solutions represent a compromise between the protection of fundamental rights and the expected benefits of AI. This has led lawmakers to respond only partially to the demands of individuals and society for the protection of their rights and freedoms, so as not to slow down the development of AI. This is even more evident in those contexts where there is no strong AI industry.

Second, also in the light of this compromise, it is crucial to provide guiding principles for the development of AI. These should not simply repeat existing legal requirements and principles, but contextualise them in the field of AI and only introduce new ones where necessary to address new challenges.

Third, the crucial role of the risk-based approach (made all the more important in the absence of detailed guiding principles) requires both a harmonised approach consistent with risk management theory – for all cases where risk is assessed – and the development of a specific methodology for the impact on fundamental rights. The latter should be based on

key criteria and variables and be properly implemented by the competent authorities and not be delegated to standardisation bodies.

At the current stage of the regulatory debate (August 2023), it is not possible to say whether all these objectives will be achieved in the AI Act, or whether they will be part of a further implementation strategy for AI laws that will pave the way for a second generation of such laws.

G. Epilogue

In revising the proofs of this chapter, it is worth noting that the final version of the AI Act does not address the two main issues mentioned in the conclusions, nor has the Council of Europe's Framework Convention on AI[83] provided a more robust and methodologically accurate response.

The AI Act, in its final version, has maintained the risk-based approach focused on high-risk categories and self-assessment, abandoning the mechanism based on a reasoned notification to the competent national supervisory authority for those systems that the AI providers do not consider to pose high risks, in favour of an explicit derogation provided for in Article 6(3) for those systems that do "not pose a significant risk of harm to the health, safety or fundamental rights of natural persons".[84]

Moreover, the general list of principles proposed by the Parliament is no longer present in the final text, but given the shortcomings mentioned above, this does not significantly change the situation and confirms the central role of the FRIA.

With regard to impact assessment, the final version of the AI Act does not solve the problem of the lack of harmonisation between the different impact assessment procedures, nor does it provide guidance on specific

83 The text of the Convention is available here: https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence.

84 According to this Article, this is the case where one of the following conditions is met: (i) the AI system is intended to perform a narrow procedural task; (ii) the AI system is intended to improve the result of a previously completed human activity; (iii) the AI system is intended to detect decision-making patterns or deviations from prior decision-making patterns and is not meant to replace or influence the previously completed human assessment, without proper human review; or (iv) the AI system is intended to perform a preparatory task to an assessment relevant for the purposes of the use cases listed in Annex III. A provider who considers that an AI system is not high risk under these conditions shall document its assessment and shall be subject to the registration obligation set out in Article 49(2) of the AI Act.

criteria and methodology for the FRIA. Worse, the trilogue weakened and made less precise the Parliament's proposal on the FRIA.

Although the European Parliament's proposal did not outline the key parameters for risk assessment, it included many elements of the FRIA and provided for a higher level of detail compared to the final text of the AI Act,[85] where some elements are implicit (e.g., the mitigation plan, the consideration of vulnerability, a clear description of the components of risk assessment).

Another important part of the European Parliament's proposal was the role of participation in risk assessment.[86] Unfortunately, as with the GDPR, the final version of the AI Act does not give due attention to participation, contrary to best practice in risk assessment.

However, the main difference between the European Parliament's proposal and the adopted text concerns the scope of the FRIA. Under pressure from the other two co-legislators, it was restricted to a limited area, whereas the text proposed by the Parliament referred to all high-risk AI systems as defined in Article 6(2), with the sole exception of systems used for management and operation of critical infrastructure. The final text maintains this exception but significantly narrows the general scope of the FRIA, which now covers only (i) deployers that are bodies governed by public law and private entities providing public services, and (ii) AI systems used to evaluate the creditworthiness of natural persons or for credit scoring (with the exception of AI systems used for the detection of financial fraud), and for risk assessment and pricing in life and health insurance.[87]

Although this narrow scope of the FRIA is less satisfactory from the perspective of the protection of fundamental rights and creates an imbalance between the general obligation for AI providers to assess the impact on fundamental rights of all high-risk AI systems in the context of the confor-

85 See Article 27 of the final version of the AI Act.

86 See Article 29a.4, AI Act EP ("In the course of the impact assessment, the deployer, with the exception of SMEs, shall notify national supervisory authority and relevant stakeholders and shall, to best extent possible, involve representatives of the persons or groups of persons that are likely to be affected by the high-risk AI system, as identified in paragraph 1, including but not limited to: equality bodies, consumer protection agencies, social partners and data protection agencies, with a view to receiving input into the impact assessment. The deployer shall allow a period of six weeks for bodies to respond. SMEs may voluntarily apply the provisions laid down in this paragraph. In the case referred to in Article 47(1), public authorities may be exempted from this obligations.").

87 See Annex III, 5 (b) and (c), AI Act.

mity assessment procedure and the specific obligation for deployers, it does not prevent the adoption of a broader use of this instrument based on the obligation to protect fundamental rights established at EU and national level, and facilitating the accountability of AI operators in this respect.

Given the nature of fundamental rights and the level of protection afforded to them by the Charter of Fundamental Rights of the European Union and national constitutional charters, this assessment must necessarily avoid any prejudice to them. This means that the FRIA cannot simply be a final check without influence on the AI design. On the contrary, potential impacts must be properly addressed on the basis of a sound methodological approach in order to meet the obligations to protect fundamental rights.

In this regard, academia can actively contribute to filling the existing gaps in the theoretical and methodological elaboration of the FRIA, as outlined in the AI Act, in order to facilitate the future work of EU and national authorities and AI operators in placing this key tool for human-centric and trustworthy AI at the heart of the EU approach to AI design and development.[88]

88 For a more detailed analysis and methodological guidelines for FRIA see Mantelero, Alessandro. «The Fundamental Rights Impact Assessment (FRIA) in the AI Act: roots, legal obligations and key elements for a model template». *Computer Law & Security Review* 54 (2024): 106020, https://doi.org/10.1016/j.clsr.2024.106020, which aims to fill existing gaps in the theoretical and methodological elaboration of the FRIA, as outlined in the AI Act, by defining the building blocks of a model template for the FRIA in a manner consistent with the rationale and scope of the AI Act.