ISSN 2281-5147

ETHICS LOST IN TRANSLATION: TRUSTWORTHY AI FROM GOVERNANCE TO REGULATION

Irina Carnat*

Abstract

The policy reaction of the European Union to the societal quest for fairer and more transparent Artificial Intelligence (AI) was immediate, yet partially unsatisfactory. Given the inadequacy of the liability regime and the absence of appropriate regulation of AI, the appointed High-Level Expert Group turned to ethics to develop the notion of Trustworthy AI. Generally, AI ethics was first tasked with providing guidance for global consensus on principles for AI governance. However, such ethics guidelines lack appropriate enforcement mechanisms, hence the unsuitability to provide clear guidance for AI regulation, leading to a potential phenomenon of 'ethics washing'. The present article unveils the quintessentially political nature of the ethics guidelines, arguing that the strictly ethical nature of the guidelines got lost in translation in light of the proposed EU regulation of AI. Instead, it is claimed that the principle of accountability bridges the gap between the ethics guidelines and the regulatory framework, as shown by a first comparative glance at the US-proposed Algorithmic Accountability Act and the EU-proposed AI Act through the lenses of the recently proposed amendments concerning foundation models, thus providing the necessary enforcement mechanism to achieve trust in AI.

Table of contents:

* Research Fellow at the Institute of Law, Politics and Development, LIDER-Lab, Sant'Anna School of Advanced Studies, and National Ph.D. Student in Artificial Intelligence for Society, University of Pisa. Irina.Carnat@santannapisa.it. This research was financially supported by the European Union's Horizon 2020 research and innovation programme under the following grant agreements: 'SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics' GA 871042 and by the project 'Biorobotics Research and Innovation Engineering Facilities "IR0000036" – CUP J13C22000400007'.

ISSN 2281-5147

Keywords

Trustworthy AI - Algorithmic accountability - AI regulation - AI governance - AI ethics

1. Introduction

The field of artificial intelligence ethics has largely emerged as a first response to the individual and societal harms that the misuse, abuse, poor design, or negative unintended consequences of AI systems may cause.[1] Absent an appropriate regulatory framework, given the inadequacy of the pre-existing product liability regime[2] to tackle the risks specifically posed by AI, the European Union recognized the need for ethics guidelines in order to create cohesion and consensus among its Member States on the issue of how to regulate artificial intelligence. These guidelines were meant to draw

[1] David Leslie, 'Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector' (2019) <https://zenodo.org/record/3240529> accessed 18 August 2022.

[2] Note from the outset that civil liability rules were already under scrutiny to assess whether fit for purpose in the context of AI. An Expert Group on liability and new technologies tasked with the evaluation of the Product Liability Directive was appointed in 2018. More on their activities available at: https://ec.europa.eu/transparency/expert-groups-register/screen/expert-groups/consult?lang=en&do=groupDetail.groupDetail&groupID=3592 accessed 25 July 2023.

ISSN 2281-5147

attention to how Europe would approach the regulation of AI with respect to other countries in the global competitive market. However, the potential for 'ethics washing'[3] or the superficial adoption of ethical standards through codes of conduct without genuine commitment made it clear that a more robust regulatory approach was necessary. If the goal is to establish *trust* in the technology to foster its development while guaranteeing the protection of individuals, not only ethical considerations are required beforehand, but also a mechanism to ensure their practical implementation.

In the first part of this paper, the ethics guidelines adopted in the EU are analyzed to unveil their quintessentially political intent. By briefly exploring both the capabilities and limitations of ethics guidelines and principles for AI governance, their shortcoming in providing clear guidance for the adoption of a regulatory framework for AI are addressed.

The second part introduces and explains the principle of accountability as a possible solution to the lack of appropriate enforcement mechanisms, bridging the gap between the ethics guidelines and the legal framework, favoring legal certainty in the context of AI development, deployment, and use.

The third part deals with the risk-based regulatory approach adopted in the proposed EU regulation on AI, to show how the allegedly ethical principles were 'translated' into technical and compliance requirements, mainly regarding high-risk AI systems. Additionally, some considerations are made concerning the US regulatory framework, as laid down in the proposed Algorithmic Accountability Act, along with the latest developments in the EU concerning the regulation of foundation models.

2. What the EU wanted: the policy statement on Trustworthy AI

Facing the challenges arising from AI, the European Union's reaction was to create a common approach[4] across the Member States to boost research and industrial

[3] Giovanni Comandé, 'Unfolding the Legal Component of Trustworthy AI: A Must to Avoid Ethics Washing' (Social Science Research Network 2020) SSRN Scholarly Paper 3690633 <https://papers.ssrn.com/abstract=3690633> accessed 25 July 2023.

[4] All the relevant documents on the European approach to Artificial Intelligence are presented in detail in a timeline available here: https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence accessed 19 April 2023.

ISSN 2281-5147

capacity while ensuring safety and fundamental rights. Such a policy statement aims to strike the right balance between innovation and protection of fundamental rights, avoiding a regulatory chilling effect on technological advancement.

The Declaration of Cooperation, signed in 2018 by all EU Member States plus Norway, was a first step towards the European approach to Artificial Intelligence.[5] It was then followed by the European Commission's Communication on Artificial Intelligence for Europe,[6] presented in April 2018 and endorsed by the European Council in June 2018, which aimed at strengthening the UE's competitive position on the international landscape with respect to the USA and China, leveraging on its world-leading AI research community to establish its leadership in AI regulation. Three main goals are pursued at this stage: 1) boosting the technological and industrial capacity and the AI uptake across the economy by increasing investments in research and innovation, along with data availability; 2) preparing for the socioeconomic changes, especially in the labor market; and 3) ensuring an appropriate ethical and legal framework.

Focusing on this last policy goal, relevant to the study at hand, the European Commission explicitly stated that for the development and use of AI, an environment of *trust and accountability* (emphasis added) is needed.[7] For this purpose, the EU could rely on the values set out in Article 2 of the TUE[8] and the Charter of Fundamental Rights,[9] as well as on a strong and established regulatory framework in terms of safety standards and product liability, network and information system security, and protection of personal data,[10] to be complemented by the forthcoming proposals under the Digital Single Market strategy. Such a policy statement was supported by

[5] Building on the progress towards the creation of a Digital Single Market, the Declaration had the primary goal to ensure "an adequate legal and ethical framework, building on EU fundamental rights and values, including privacy and protection of personal data, as well as principles such as transparency and accountability". See European Commission, Declaration of cooperation on Artificial Intelligence (2018).
[6] European Commission, Communication on Artificial Intelligence for Europe, (COM(2018) 237 final).
[7] Ibidem.
[8] Article 2 of the Treaty on European Union states: "The Union is founded on the values of respect for human dignity, freedom, democracy, equality, the rule of law and respect for human rights, including the rights of persons belonging to minorities. These values are common to the Member States in a society in which pluralism, non-discrimination, tolerance, justice, solidarity and equality between women and men prevail."
[9] Charter of Fundamental Rights of the European Union (2012).
[10] The Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) was bound to enter into force a month later, in May 2018.

ISSN 2281-5147

the Coordinated Plan on Artificial Intelligence[11], which stressed the need for ethical guidelines to build societal trust in AI, unveiling the EU's ambition to become the global leader in AI governance. It is worth noticing that the initiative to develop ethics guidelines is coupled with an extensive review of the pre-existing liability regime, the Product Liability Directive in particular, to ensure that they are fit for purpose considering new risks posed by emerging digital technologies.[12]

On these grounds, after launching the European AI Alliance, in June 2018 the Commission established the High-Level Expert Group on Artificial Intelligence (hereinafter referred to as 'HLEG'), an independent group of 52 members mandated with the drafting of two deliverables: (1) AI Ethics Guidelines (hereinafter referred to as 'Ethics Guidelines'), and (2) Policy and Investment Recommendations (hereinafter 'Recommendations').[13]

Before delving deeper into the merits of the Ethics Guidelines as developed by the HLEG, three preliminary remarks are worth mentioning.

First, the role of ethics in AI governance is not new. In fact, there is an extensive study on how governments and private organizations adopted a set of seemingly ethical principles and guidance for AI governance,[14] with a specific goal in mind: building a global consensus on the need to regulate, first, and on the standards for such regulation, second.[15] As such, founded on rather abstract and somewhat vague principles that deliberately leave room for interpretation, the field of AI ethics has played a preliminary role in AI governance, by making governments willing to actively

[11] European Commission, Coordinated Plan on Artificial Intelligence (COM(2018) 795 final.
[12] See European Commission, Staff Working Document, Liability for Emerging Digital Technologies, SWD(2018) 137 final.
[13] While the Guidelines were meant to provide guidance to individuals or organizations that develop, deploy, or otherwise use AI systems, the Recommendations addressed European institutions and Member States for their policy strategies on AI. See High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI (April 2019).
[14] See, for instance, Lewin Schmitt, 'Mapping Global AI Governance: A Nascent Regime in a Fragmented Landscape' (2022) 2 AI and Ethics 303 <https://doi.org/10.1007/s43681-021-00083-y> accessed 16 May 2023; Urs Gasser and Virgilio AF Almeida, 'A Layered Model for AI Governance' (2017) 21 IEEE Internet Computing 58. See also James Butcher and Irakli Beridze, 'What Is the State of Artificial Intelligence Governance Globally?' (2019) 164 The RUSI Journal 88. See also Daniel S Schiff and others, 'What's Next for AI Ethics, Policy, and Governance? A Global Overview' (SocArXiv, 17 December 2019) <https://osf.io/preprints/socarxiv/8jaz4/> accessed 3 May 2023.
[15] Virginia Dignum, 'Ethics in Artificial Intelligence: Introduction to the Special Issue' (2018) 20 Ethics and Information Technology 1 <https://doi.org/10.1007/s10676-018-9450-z> accessed 8 May 2023.

ISSN 2281-5147

engage in achieving a common approach for AI,[16] despite inevitable divergences due to cultural and political contexts.[17]

Second, despite the existence of an extensive body of ethical principles and guidelines to be applied to AI at the time the HLEG was appointed,[18] in the European Commission's view such a proliferation of guidelines may prove to be an obstacle to the development of the European Single Market, hence the need for a truly European approach to AI.[19] This would also reinforce the EU's ambition to become a global leader in cutting-edge AI by building individual and collective trust in the technology. Moreover, the European AI Alliance was contextually established: an open multi-stakeholder platform providing input for the HLEG's tasks.[20] It is clear thus how ethics plays a central role in the European Commission's strategy aimed at enhancing the democratization process of AI regulation through a bottom-up approach of stakeholders' consultation, while at the same time strengthening the EU's global position as the pioneer of AI regulation.

Third, in comparison to the United States[21] and China,[22] which are both major players in the global AI market, focusing mainly on the utility and cost-benefit analysis,[23] the

[16] Schmitt (n 14).
[17] Seán S ÓhÉigeartaigh and others, 'Overcoming Barriers to Cross-Cultural Cooperation in AI Ethics and Governance' (2020) 33 Philosophy & Technology 571 <https://doi.org/10.1007/s13347-020-00402-x> accessed 3 May 2023.
[18] For an extensive inventory of principles, voluntary commitments, and frameworks, see the Algorithm Watch in 'AI Ethics Guidelines Global Inventory', available at: https://inventory.algorithmwatch.org/ accessed 19 December 2022. See also Anna Jobin, Marcello Ienca and Effy Vayena, 'The Global Landscape of AI Ethics Guidelines' (2019) 1 Nature Machine Intelligence 389 <https://www.nature.com/articles/s42256-019-0088-2> accessed 16 May 2022.
[19] European Commission, Communication on Artificial Intelligence for Europe, (COM(2018) 237 final.
[20] Since its establishment in 2018, the European AI Alliance assembly has engaged more than 6.000 stakeholders. More on its activities and assemblies can be found here: https://ec.europa.eu/futurium/en/european-ai-alliance/ai-hleg-steering-group-european-ai-alliance.html accessed 25 July 2023.
[21] The US national strategy on AI, while still referring to the trustworthy dimension of the technology, mainly focuses on establishing leadership in research and development for AI's integration across all sectors of economy and society, such as national defense: https://www.ai.gov/legislation-and-executive-orders/ accessed 19 April 2023. Such an approach is confirmed by the latest released report by the National Security Commission on Artificial Intelligence in 2021 https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf accessed 19 April 2023.
[22] Huw Roberts and others, 'The Chinese Approach to Artificial Intelligence: An Analysis of Policy, Ethics, and Regulation' (2021) 36 AI & SOCIETY 59 <https://doi.org/10.1007/s00146-020-00992-2> accessed 23 January 2023.
[23] National Science Technology Council (2019) The national artificial intelligence research and development strategic plan: 2019 update. National Science and Technology Council (US)-Committee on Artificial Intelligence. Retrieved October 01, 2021, from https://www.hsdl.org/?abstract&did=831483

ISSN 2281-5147

EU has taken a more proactive approach to regulating the development and use of AI.[24] This approach sets the EU apart from other countries, which may be more focused on the economic benefits of AI and less concerned with its purely ethical implications. As a result, the EU's regulatory leadership on AI, rooted in ethical principles, could help to shape the global conversation on this technology[25] and set the standard for other countries to follow.[26] The very same trustworthy dimension of AI founded in ethics-like principles is promoted at the level of Organisation for Economic Co-operation and Development ('OECD'), listing both values-based principles, such as fairness, transparency, explainability, robustness, accountability, etc., as well as recommendations for policymakers.[27] Likewise, the United States has issued the Blueprint for an AI Bill of Rights,[28] which, despite different phrasing, aims at setting the desirable principles for trustworthy AI, namely, safety, fairness and non-discrimination, explainability, privacy and data protection, and human oversight.[29] Generally, scholars have identified overlapping and recurring principles towards which global consensus is deemed reached.[30] Although a thorough analysis of all

[24] Giovanni De Gregorio, 'The Rise of Digital Constitutionalism in the European Union' (2021) 19 International Journal of Constitutional Law 41 <https://doi.org/10.1093/icon/moab001> accessed 23 January 2023.

[25] Elettra Bietti, 'From Ethics Washing to Ethics Bashing: A View on Tech Ethics from within Moral Philosophy', *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2020) <https://dl.acm.org/doi/10.1145/3351095.3372860> accessed 8 June 2023.

[26] Nathalie A Smuha, 'The EU Approach to Ethics Guidelines for Trustworthy Artificial Intelligence' (Social Science Research Network 2019) SSRN Scholarly Paper 3443537 <https://papers.ssrn.com/abstract=3443537> accessed 16 May 2022.

[27] See OECD, AI Principles overview, available at: https://oecd.ai/en/ai-principles accessed 25 July 2023.

[28] The White House, *Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People*, October 2022 https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf accessed 31 January 2023.

[29] It is worth noting that the Blueprint on an AI Bill of Rights is consistent with the previous US Executive Order 13960 on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government, which requires that certain federal agencies comply with nine stated principles when designing, developing, acquiring or using AI for purposes other than national security or defense. See The White House, Executive Order of the President n. 13960 of 3 December 2020: https://www.govinfo.gov/content/pkg/FR-2020-12-08/pdf/2020-27065.pdf accessed 7 June 2023.

[30] See, e.g., Daniel Schiff and others, 'AI Ethics in the Public, Private, and NGO Sectors: A Review of a Global Document Collection' (2021), 2, 31. The study analyzed 112 documents from 25 countries produced between 2016 and mid-2019, revealing significant variations in ethical coverage and depth across public, private, and non-governmental organizations, with NGO and public sector documents demonstrating broader ethical coverage, greater engagement with law and regulation, and more participatory processes, highlighting differing beliefs about organizational responsibilities, the role of experts versus public representatives, and the balance between prosocial and economic objectives. See also Luciano Floridi and Josh Cowls, 'A Unified Framework of Five Principles for AI in Society' (2019) 1 Harvard Data Science Review <https://hdsr.mitpress.mit.edu/pub/l0jsh9d1/release/8> accessed 7 June 2023. The study shows how four

ISSN 2281-5147

principles would go beyond the scope of the present article, it is worth highlighting that, according to one such taxonomy, five topics are more prominent: social responsibility, transparency, fairness, privacy, and finally safety and reliability.[31]

Against this background, it is clear how the HLEG's work was meant to facilitate the global consensus on such principles grounded in traditional democratic values,[32] thus strengthening the EU's leadership on AI governance.

2.1. What the HLEG delivered: the Ethics Guidelines for Trustworthy AI

According to the Ethics Guidelines, for AI to be trustworthy it must be: 1) lawful, 2) ethical, and 3) robust, both from a technical and social perspective. The framework provided by the HLEG consists of three chapters, briefly explained below:

Chapter I - Foundations of Trustworthy AI, laying out a fundamental rights-based approach, identifying and describing the relevant ethical principles. Such principles are: i) respect for human autonomy, based on a human-centric approach and human oversight; ii) prevention of harm, making sure that AI systems are safe and secure, with particular regard to vulnerable persons; iii) fairness, with a view to ensuring equal and just distribution of both benefits and costs without unfair bias, discrimination, and stigmatization; iv) explicability, concerning the so-called 'black-box' problem, aiming at ensuring that AI systems are sufficiently transparent, traceable, and auditable.

Chapter II - Realizing Trustworthy AI, listing seven key requirements to be implemented and met throughout the life cycle of an AI system, adopting *technical and non-technical* (emphasis added) methods. Such key requirements are: human agency and oversight; technical robustness and safety; privacy and data governance, transparency;

---

out of the five most recurrent principles, namely beneficence, non-maleficence, autonomy, and justice, stem from traditional bioethics principles, to which explicability is added specifically from the field of AI.

[31] Schiff and others (n 30), p. 8. Similarly, the principles of transparency, justice and fairness, non-maleficence, responsibility, and privacy can be found in more than half of the evaluated guidelines, according to Jobin, Ienca and Vayena (n 18). See also Eva Erman and Markus Furendal, 'The Global Governance of Artificial Intelligence: Some Normative Concerns' (2022) Moral Philosophy and Politics <https://www.degruyter.com/document/doi/10.1515/mopp-2020-0046/html?lang=en> accessed 16 May 2022, p. 7 ff.

[32] Erman and Furendal, (n 31), p. 8.

ISSN 2281-5147

diversity, non-discrimination and fairness; societal and environmental wellbeing; and accountability.

Chapter III - Assessing Trustworthy AI: it sets up an Assessment List for the Trustworthiness of the AI system (hereinafter 'ALTAI').[33] The goal of such a non-exhaustive, pilot version, assessment list is to operationalize the concept of Trustworthy AI, with the caveat that compliance thereof does not prove compliance with legal requirements, nor is it intended as guidance for such compliance. Instead, it "*encourages reflection on how Trustworthy AI can be operationalized, and on the potential steps that should be taken in this regard*".[34] Therefore, how to make such principles of Trustworthy AI operationalizable in practice remains an open question.

The European Commission contextually released a Communication in support of the Ethics Guidelines, stating that it should be considered valuable input for the forthcoming policymaking.[35] However, besides reaffirming that building trust is a prerequisite for the human approach to AI, it also restated the goal to play a leadership role in the development of international AI guidelines and potentially a related assessment mechanism. For this purpose, the EU plans to strengthen cooperation with like-minded countries, engage in dialogues with non-EU countries to build a consensus on human-centric AI, and to participate in the development of relevant international standards to promote its vision of human-centric AI.

2.2. The unsuitable role of the Ethics Guidelines for AI regulation

Having clarified how the current field of AI ethics developed in an attempt to strengthen the consensus on governance strategies for AI on a global scale, it is worth noticing that all the ethics guidelines suffer from the same shortcoming when applied as regulatory tools.[36] A common criticism is the lack of clarity due to overlapping and

[33] https://futurium.ec.europa.eu/en/european-ai-alliance/pages/welcome-altai-portal accessed 4 January 2022. Following a pilot process with more than 350 stakeholders, the ALTAI was revised, and a prototype web tool was released.
[34] European Commission, Content and Technology Directorate General for Communications Networks, and High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI* (2019) <https://data.europa.eu/doi/10.2759/346720> accessed 13 May 2022. p. 24.
[35] European Commission, Building Trust in Human-Centric Artificial Intelligence (COM(2019) 168 final), p. 4.
[36] For an extensive evaluation of the most relevant ethics guidelines, see Thilo Hagendorff, 'The Ethics of AI Ethics: An Evaluation of Guidelines' (2020) 30 Minds and Machines 99 <http://link.springer.com/10.1007/s11023-020-09517-8> accessed 16 May 2022.

ISSN 2281-5147

ambiguous terminology of ethical principles.[37] Another critical issue is that ethical guidelines do not offer methods for solving tensions and conflicts between principles, which should account for appropriate societal stakeholder participation.[38] Other criticisms derive from the circumstance that ethics in the early stages of AI regulation has been associated with a 'soft governance' approach: scholars have pointed out that without common criteria to evaluate the quality of ethical and legal commitments around the impact of AI on fundamental rights, there is a considerable danger for such frameworks to become arbitrary, optional or meaningless rather than substantive, effective and rigorous ways to design technologies.[39]

Although the so-called 'soft law' approach might cause companies to adopt better behavior models and operate more ethically with respect to commercial, social, and environmental values, it may result in a shift in legal responsibility if not finally reflected in legally binding rules.[40] In fact, private organizations adopting and formally implementing codes of conduct are still subject to a form of 'self-regulation' that lacks proper implementation[41], leading thus to the phenomenon of 'ethics washing'.[42]

Concerning the HLEG, although some of the appointed members were lawyers, law professors, or law graduates,[43] it expressly did not deal with the lawfulness component.[44] Nonetheless, the three components of Trustworthy AI are clearly interconnected, and the overall approach is grounded on fundamental rights. Several attempts to make ethics guidelines actionable and useful to stimulate policymaking

[37] Jess Whittlestone and others, 'Ethical and Societal Implications of Algorithms, Data, and Artificial Intelligence: A Roadmap for Research', London: Nuffield Foundation (2019).
[38] Hagendorff (n 36). See also Jess Whittlestone and others, 'The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions', *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (ACM 2019) <https://dl.acm.org/doi/10.1145/3306618.3314289> accessed 16 May 2022.
[39] Brent Mittelstadt, 'Principles Alone Cannot Guarantee Ethical AI' (2019) 1 Nature Machine Intelligence 501 <https://www.nature.com/articles/s42256-019-0114-4> accessed 16 May 2022.
[40] Hagendorff (n 36).
[41] Stefan Larsson, 'On the Governance of Artificial Intelligence through Ethics Guidelines' (2020) 7 Asian Journal of Law and Society 437 <https://www.cambridge.org/core/journals/asian-journal-of-law-and-society/article/on-the-governance-of-artificial-intelligence-through-ethics-guidelines/992BD33CA7CBBE83E2FBBF6B0179896C> accessed 3 May 2023.
[42] Ben Wagner, 'Ethics As An Escape From Regulation. From "Ethics-Washing" To Ethics-Shopping?', *Ethics As An Escape From Regulation. From "Ethics-Washing" To Ethics-Shopping?* (Amsterdam University Press 2018) <https://www.degruyter.com/document/doi/10.1515/9789048550180-016/html?lang=en> accessed 18 May 2022.
[43] The composition and expertise of each member of the High-Level Expert Group can be found here: https://ec.europa.eu/futurium/en/european-ai-alliance/ai-hleg-steering-group-european-ai-alliance.html accessed 19 April 2023.
[44] Giovanni Comandé (n 3).

ISSN 2281-5147

were proposed,[45] yet the most relevant issue is the lack of enforcement mechanisms that guarantee that they are adhered to, monitored, and regulated.[46] A recent example is provided with respect to an end-to-end AI system deployed for verifying the severity of lung damage caused by COVID-19, where the HLEG's Ethics Guidelines were used to assess the trustworthiness of the AI system.[47] The use case proves how the lack of clear guidance on how to apply the principles and the key requirements led to unclear results in assessing the trustworthiness of an AI system due to difficulties in mapping the relevant ethical principles and solving the contrasts among them.[48] Moreover, given the highly multidisciplinary nature of the assessing group,[49] mapping the issues and the corresponding key requirements for Trustworthy AI was very challenging, which may have caused vague and unclear results.[50] This is but one example how difficult the work of multidisciplinary auditing groups may be, if harmonized standards are not available for specific, oftentimes highly sectorial, applications of the AI system.

With specific regard to the public sector, multiple and diverse attempts to incorporate ethics-by-design have been developed. For instance, the Alan Turing Institute in 2019 envisaged a responsible AI project delivery ecosystem based on three building blocks, which are respectively composed of other value-based or procedural principles.[51] In this case, the partial divergence from the HLEG's Ethics Guidelines demonstrates how a purely ethics-based approach to AI may lead to arbitrary frameworks that may be either very complex to navigate or of little practical usefulness if not

[45] Charlotte Stix, 'Actionable Principles for Artificial Intelligence Policy: Three Pathways' (2021) 27 Science and Engineering Ethics 15. Luciano Floridi and others, 'AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations' (2018) 28 Minds and Machines 689 <https://doi.org/10.1007/s11023-018-9482-5> accessed 16 May 2022.

[46] Connor Rees and Berndt Müller, 'All That Glitters Is Not Gold: Trustworthy and Ethical AI Principles' (2022) AI and ethics, p. 11.

[47] Himanshi Allahabadi and others, 'Assessing Trustworthy AI in Times of COVID-19. Deep Learning for Predicting a Multi-Regional Score Conveying the Degree of Lung Compromise in COVID-19 Patients' (2022) IEEE Transactions on Technology and Society, 1.

[48] Ibidem.

[49] The group was composed by 58 members split in 8 working groups based on their field of expertise, namely technical, ethics, healthcare, law and social sciences.

[50] "*Both the mappings and the consolidation of the mappings involve subjective decision-making components*", Allahabadi and others (n 47), p. 15.

[51] Leslie (n 1). In particular, the first building block consists of ethical values, namely the SUM Values (Support, Underwrite, and Motivate), respectively composed of four key notions, notably Respect, Connect, Care, and Protect; the second building block consists of actionable principles (FAST Track Principles): Fairness, Accountability, Sustainability, and Transparency; finally, the third building block requires a process-based governance framework (PBG Framework), which operationalises the first two building blocks.

ISSN 2281-5147

operationalized through an enforcement mechanism[52] grounded in legally binding rules. Moreover, the need for a more detailed risk management plan identifying the accountable subjects for the AI system's failures has been advanced, showing how the lack of a clear definition of roles and responsibilities under a liability framework hinders the full potential of an AI system.[53]

Even from a philosophical perspective, the functional use of ethics as a means to achieve targets of trustworthiness of a system to increase its social acceptability is problematic in the sense that it sees the desirability of AI as an axiom[54]. The concept of trust, when applied to AI, is somewhat problematic.[55] Simply put, we don't trust technology, we trust people. In the normative account, the trustee must be held responsible for its actions - which AI per se cannot be -; on the contrary, for AI to be reliable, the burden of responsibility must be necessarily placed on those developing, deploying, and using these technologies.[56]

For all the reasons expressed above, the real problem that should have been defined by the HLEG is how to allocate responsibility for decisions concerning the development and deployment of AI. In fact, it is explicitly stated in the European Commission's White Paper, published in December 2020[57] consequent to the feedback on - and inevitable criticism around - the HLEG's Ethics Guidelines and Recommendations, that the actual challenge to fundamental rights and safety from a liability perspective in the context of AI is posed by the difficulty of tracing back potentially harmful AI-based decisions that may have caused damage. One key result of the feedback process that the European Commission took into consideration is that some of the requirements in the ethics guidelines are already covered by existing

[52] Jobin, Ienca and Vayena (n 18).
[53] Allahabadi and others (n 47), p. 11.
[54] Bernd Carsten Stahl, 'From Computer Ethics and the Ethics of AI towards an Ethics of Digital Ecosystems' (2022) 2 AI and Ethics 65 <https://doi.org/10.1007/s43681-021-00080-1> accessed 16 May 2022.
[55] Matthias Braun, Hannah Bleher and Patrik Hummel, 'A Leap of Faith: Is There a Formula for "Trustworthy" AI?' (2021) 51 Hastings Center Report 17 <https://onlinelibrary.wiley.com/doi/abs/10.1002/hast.1207> accessed 13 May 2022.
[56] An extensive study on the concept of trust in AI is provided by Mark Ryan, 'In AI We Trust: Ethics, Artificial Intelligence, and Reliability' (2020) 26 Science and Engineering Ethics 2749 <https://doi.org/10.1007/s11948-020-00228-y> accessed 16 May 2022.
[57] European Commission, White Paper on Artificial Intelligence – A European approach to excellence and trust (COM(2020) 65 final.

ISSN 2281-5147

regulatory regimes,[58] while others – perhaps the most relevant ones in the context of AI, such as transparency, traceability, and human oversight – are not reflected in existing legislation, thus the need to address them in a more attentive and specific manner.[59] As such, one fundamental failure of the AI ethics guidelines in the absence of clear identification of the responsible party for the enforcement, and measurement of suggested measures for AI systems, as well as the importance of accountability for the successful implementation of principles developed by different organizations.[60]

However obvious it may seem, ethical principles alone do not guarantee compliance. Without a fundamental shift in regulation, the translation of principles into practice it will remain a competitive, not cooperative, process. Therefore, the elephant in the room of AI ethics is how to fit it under an umbrella of (legal) enforceability.

To ensure that ethical considerations are taken seriously and effectively integrated into decision-making processes, they should at minimum conform to certain basic criteria, including external participation and engagement with all relevant stakeholders; mechanisms for external independent oversight; transparent decision-making procedures that clearly explain why certain decisions were taken; a stable set of standards that can be plausibly justified and that do not arbitrarily prioritize certain values, principles, or rights over others; a commitment to not substituting fundamental rights or human rights; and a clear statement on the relationship between ethical commitments and existing legal or regulatory frameworks, particularly in cases where these frameworks may be in conflict.[61]

2.3. A partial step forward: the Policy and Investment Recommendation

As a closure to this 'ethical' framework, three months after the ethics guidelines, the HLEG also published the Recommendations, providing policies and measures that

[58] First and foremost, the General Data Protection Regulation. See also the recently adopted 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), 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).
[59] European Parliament. Directorate General for Parliamentary Research Services., *A Governance Framework for Algorithmic Accountability and Transparency*. (Publications Office 2019) <https://data.europa.eu/doi/10.2861/59990> accessed 5 May 2023.
[60] Rees and Müller (n 46), pp. 10-11.
[61] Wagner (n 42), p. 3.

ISSN 2281-5147

European institutions and Member States should adopt to ensure the beneficial impact of AI, both by promoting its development and competitiveness in Europe and by protecting individuals and society from the risks it poses. While being drafted by the same members that laid down the Ethics Guidelines, the Recommendations have an exquisite political nature. In fact, they are addressed to EU institutions and Member States to stimulate policymaking and investments in sectors that are crucial for the EU economy in a globally competitive landscape. The final goal is, once again, to strengthen the European Single Market.[62]

However, perhaps the most important yet overlooked aspect is that the Recommendations (finally) deal with the 'lawful' component that was deliberately left undealt within the Ethics Guidelines. Here the HLEG acknowledges that Trustworthy AI requires appropriate governance and, while praising the EU's sound regulatory expertise, it also proposes several recommendations specifically for AI.[63] First, it suggests a risk-based approach based on proportionality and the precautionary principle, according to which certain AI applications should be prohibited if the risk level is deemed unacceptable. Second, it calls for an evaluation and potential revision of EU laws in critical domains, such as data protection, consumer protection, non-discrimination laws, cybersecurity, competition, etc. Third, it points out the possible need for new regulation to better deal with the critical concerns brought by AI that may not be properly addressed by existing legislation. Fourth, it questions the competencies, capacities, resources, and enforcement powers of the EU institutions to face the need for stakeholder consultation, auditing and effective redress mechanisms. Finally, it urges to establish governance mechanisms for a European Single Market for Trustworthy AI through harmonized regulation and cooperation across Member States, certification and standardization measures, and a sound enforcement mechanism.

Like the Ethics Guidelines, the Recommendations were not exempt from criticism either, pointing at their lack of problem structuring and definition, of democratic legitimacy with respect to civil society and other stakeholder consultation,

[62] High-Level Expert Group on Artificial Intelligence, Policy and Investment Recommendations for Trustworthy AI, (2019), p. 7: "This more holistic vision lends itself to the creation of a European Single Market for Trustworthy AI, where Europe is in an exceptional position to put tailored policy and investment measures in place that can enable it to seize the benefits and capture the value of AI, while minimising and preventing its risks."
[63] Ibid. p. 39.

ISSN 2281-5147

infrastructure from a data protection perspective, and decisiveness in their policy recommendations.[64]

However, what is relevant for the study at hand is that while the Ethics Guidelines alone were not sufficient nor adequate to provide concrete guidance for AI regulation, the same can be said about the Recommendations but only if viewed from the 'ethics' perspective. For this reason, it is fundamental to understand the role the HLEG's deliverables played in the broader political context, where the Recommendations were an early stage but perhaps more fitting approach and a step in the right direction. In fact, the Recommendations called for *traceability* and *reporting* requirements for AI applications with a view to facilitating *auditability*, *ex-ante oversight* and *monitoring* (emphasis added) by the competent authorities, and meaningful human intervention and oversight over the deployment of AI systems in critical decision-making contexts that have an impact on safety and fundamental rights.

3. What the EU actually needed: from ethics to accountability

In acknowledging their non-binding nature, the European Commission puts the HLEG's Ethics Guidelines on a par with the President's political guidelines,[65] for the ultimate policy statement on the European approach to AI contained in the White Paper. As such, the formally 'ethical' principles contained thereof are not to be considered as a matter of ethics in the strict sense, but a mere policy statement to stimulate a legally enforceable reaction. The European Commission calls thus for a clear regulatory framework that favors both consumers with adequate protection and businesses with legal certainty. It acknowledges that only a clear European regulatory framework can build the necessary trust in the technology. To this purpose, the White Paper contains recommendations for the adoption of risk-based regulation, focusing on risks to fundamental rights and safety, considering the specific characteristics of AI-driven technologies. The adequacy of the pre-existing legal framework to provide protection was challenged by AI's opacity, complexity, unpredictability and partially

[64] Michael Veale, 'A Critical Take on the Policy Recommendations of the EU High-Level Expert Group on Artificial Intelligence' (Social Science Research Network 2020) SSRN Scholarly Paper 3475449 <https://papers.ssrn.com/abstract=3475449> accessed 16 May 2022.

[65] It also worth noting that the White Paper was released shortly after the new presidency of the European Commission assumed office at the end of 2019, under the guidance of the president Ursula von der Leyen. See President Von der Leyen, A Union that strives for more – My agenda for Europe, page 17.

ISSN 2281-5147

autonomous behavior that makes it difficult to verify compliance and guarantee effective enforcement.[66] For this reason, the pressing concern for the European Commission was to adopt the appropriate adjustments to existing EU laws to tackle the specific risks posed by AI systems.[67]

However, the Ethics Guidelines, although formally considered, were only partially reflected in the key features identified as the requirements to ensure legal certainty. In fact, the White Paper turned to features that do not recall any 'ethical' characteristic: training data; data and record-keeping; information to be provided; robustness and accuracy; human oversight. Instead, their shared normative semantic is that of a compliance and enforcement mechanism that ensures monitoring, verification, and documentation of the AI system's development process. As a matter of fact, such requirements are addressed to the actors who are best placed to address the potential risks, clearly reflecting the risk-based logic behind them.[68] According to this - revised - EU policy statement, the future regulatory framework for Trustworthy AI does not turn to the ethical discourse to draw its regulatory guideline but rather anchors it in the well-established strict liability regime[69] with a fundamental, yet often overlooked, corrective action: *accountability*. After all, since its first policy statement, the European Commission stated that "*An environment of trust and accountability around the development and use of AI is needed*" to ensure an appropriate ethical and legal framework.[70]

---

[66] European Commission (n 57).

[67] The report on the safety and liability implications of Artificial Intelligence, Internet of Things and robotics accompanying the White Paper identified the new risks presented by the emerging digital technologies and proposed specific provisions to address them. See European Commission, 'Report on the safety and liability implications of AI, the Internet of Things and Robotics' COM (2020) 64 final, 12 (Safety and Liability Report).

[68] See generally Guido Calabresi, 'The Cost of Accidents: A Legal and Economic Analysis', Yale University Press, 1970.

[69] While the HLEG reaffirms the importance of having adequate compensation schemes for damaged caused by AI, along with a mandatory insurance provision, new liability rules applicable to AI have recently been proposed: Proposal for a Directive of the European Parliament and of the Council on liability for defective products, COM(2022) 495 final; Proposal for a directive of the European Parliament and of the Council on adapting noncontractual civil liability rules to artificial intelligence (AI Liability Directive), COM(2022) 496 final. These proposed directives are based on the European Parliament's 'Resolution of 20 October 2020 with recommendations to the Commission on a civil liability regime for artificial intelligence' 2020/2014(INL) (Resolution on Civil Liability or Parliamentary Resolution) and the White Paper, European Commission (n 57). See, e.g. Gerhard Wagner, 'Robot Liability' in Sebastian Lohsse, Rainer Schulze and Dirk Staudenmayer (eds), Liability for Artificial Intelligence and the Internet of Things (Hart 2019). See also Beatrice Schütte, Lotta Majewski and Katri Havu, 'Damages Liability for Harm Caused by Artificial Intelligence – EU Law in Flux' (2021) SSRN Electronic Journal.

[70] European Commission (n 6), p. 13

ISSN 2281-5147

3.1. Conceptual foundations of the (overlooked) principle of accountability

Generally, accountability is a multifaceted concept that can adapt to a wide range of meanings and applications.[71] The most adopted conceptual framework is provided by Bovens, according to which accountability is a relational concept between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgment, and the actor may face consequences.[72] From such a general conceptual framework one can tailor a definition of accountability that reflects beliefs about the ideal relationship between actors.[73] As such, accountability is highly contextual[74] and can assume multiple forms based on the normative logic,[75] power relation,[76] or the adopted substantive notion.[77]

These notions are tailored and adapted to the specific risks and characteristics in the field of AI,[78] such as the 'black box' problem,[79] autonomous behavior, bias and

[71] Stephen Keith McGrath and Stephen Jonathan Whitty, 'Accountability and Responsibility Defined' (2018) 11 International Journal of Managing Projects in Business 687 <https://doi.org/10.1108/IJMPB-06-2017-0058> accessed 18 December 2022.

[72] Mark Bovens, 'Analysing and Assessing Accountability: A Conceptual Framework' (2007) 13 European Law Journal 447 <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0386.2007.00378.x> accessed 12 August 2022.

[73] Jonathan GS Koppell, 'Pathologies of Accountability: ICANN and the Challenge of "Multiple Accountabilities Disorder"' (2005) 65 Public Administration Review 94 <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6210.2005.00434.x> accessed 8 January 2023.

[74] Jennifer Cobbe, Michelle Seng Ah Lee and Jatinder Singh, 'Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems', *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2021) <https://dl.acm.org/doi/10.1145/3442188.3445921> accessed 24 November 2022.

[75] Joseph Donia, 'Normative Logics of Algorithmic Accountability', *2022 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2022) <https://dl.acm.org/doi/10.1145/3531146.3533123> accessed 15 November 2022.

[76] Based on the nature of the power relation which exists between the actor and the forum, accountability can be vertical, horizontal, or diagonal. See Bovens (n 72).

[77] Koppell (n 73).

[78] For an extensive study of the risks posed by AI and possible regulatory approaches, see Margot E Kaminski, 'Regulating the Risks of AI' (2022) SSRN Electronic Journal <https://www.ssrn.com/abstract=4195066> accessed 12 September 2022.

[79] See, e.g., Madalina Busuioc, Deirdre Curtin and Marco Almada, 'Reclaiming Transparency: Contesting the Logics of Secrecy within the AI Act' (2022) European Law Open 1 <https://www.cambridge.org/core/journals/european-law-open/article/reclaiming-transparency-contesting-the-logics-of-secrecy-within-the-ai-act/01B90DB4D042204EED7C4EEF6EEBE7EA> accessed 23 December 2022. See Madalina Busuioc, 'Accountable Artificial Intelligence: Holding Algorithms to Account' (2021) 81 Public Administration Review 825 <https://onlinelibrary.wiley.com/doi/abs/10.1111/puar.13293> accessed 10 August 2022.

ISSN 2281-5147

possible discrimination, etc. to develop a specific algorithmic accountability framework. To this purpose, scholars have identified the barriers to accountability in the context of AI[80] and proposed several possible solutions to address them. For instance, based on the specific regulatory goals, two different uses of accountability in AI: proactive and reactive, complementary one to another.[81] The former is implemented ex-ante, serving the purposes of compliance and oversight. The latter is implemented ex-post, serving the purposes of reporting and enforcement. Accountability tools in AI range from reviewability,[82] auditing,[83] impact assessments,[84] or other technical solutions.[85]

A thorough analysis of possible accountability solutions for AI systems would go beyond the scope and purpose of this article; for now, it suffices to observe that the solutions proposed are based on the need for more fairness, transparency, and

[80] For instance, the four barriers to accountability identified by Nissenbaum in are: 1) the problem of 'many hands'; 2) 'bugs'; 3) computers as scapegoats; and 4) ownership without liability. See Helen Nissenbaum, 'Accountability in a Computerized Society' (1996) 2 Science and Engineering Ethics 25 <http://link.springer.com/10.1007/BF02639315> accessed 20 April 2023. See also A Feder Cooper and others, 'Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning' (2022) <http://arxiv.org/abs/2202.05338> accessed 31 May 2022.

[81] See, generally, Claudio Novelli, Mariarosaria Taddeo and Luciano Floridi, 'Accountability in Artificial Intelligence: What It Is and How It Works' (3 August 2022) <https://papers.ssrn.com/abstract=4180366> accessed 6 August 2022.

[82] Chris Norval, Jennifer Cobbe and Jatinder Singh, 'Towards an Accountable Internet of Things: A Call for Reviewability' in Andrew Crabtree, Hamed Haddadi and Richard Mortier (eds), *Privacy by Design for the Internet of Things: Building accountability and security* (Institution of Engineering and Technology 2021) <https://digital-library.theiet.org/content/books/10.1049/pbse014e_ch5> accessed 8 January 2023. See also Joshua A Kroll, 'Outlining Traceability: A Principle for Operationalizing Accountability in Computing Systems', *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021) <http://arxiv.org/abs/2101.09385> accessed 16 November 2022.

[83] Inioluwa Deborah Raji, 'From Algorithmic Audits to Actual Accountability: Overcoming Practical Roadblocks on the Path to Meaningful Audit Interventions for AI Governance', *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Association for Computing Machinery 2022) <https://doi.org/10.1145/3514094.3539566> accessed 21 April 2023. See also Inioluwa Deborah Raji and others, 'Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing' (arXiv, 3 January 2020) <http://arxiv.org/abs/2001.00973> accessed 16 November 2022.

[84] See, generally, Margot E Kaminski, 'Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability' (2019) SSRN Electronic Journal <https://www.ssrn.com/abstract=3351404> accessed 12 December 2022. See also Heleen Janssen, Michelle Seng Ah Lee and Jatinder Singh, 'Practical Fundamental Rights Impact Assessments' (2022) 30 International Journal of Law and Information Technology 200 <https://academic.oup.com/ijlit/article/30/2/200/6835507> accessed 24 November 2022. See also Jacob Metcalf and others, 'Algorithmic Impact Assessments and Accountability: The Co-Construction of Impacts', *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2021) <https://doi.org/10.1145/3442188.3445935> accessed 16 November 2022.

[85] Timnit Gebru and others, 'Datasheets for Datasets' (arXiv, 1 December 2021) <http://arxiv.org/abs/1803.09010> accessed 27 March 2023.

ISSN 2281-5147

explainability,[86] but do not share an 'ethical' nature. Instead, they are addressed from an accountability viewpoint resulting in requirements for AI developers to meet,[87] while at the same time providing the justification for why certain decisions were adopted in the first place.[88] For instance, transparency is often associated with the need to scrutinize the inner workings of a system to understand the consequences of its operation.[89] However, transparency alone lacks practical effectiveness unless an additional enforceable framework is established to enable principals to hold agents accountable using transparent explanations and justifications.[90] In this sense, transparency is instrumental in achieving accountability, holding the system's creator, provider or operator responsible for potentially harmful outcomes.[91] A similar remark can be made with regard to explainability, the demand of which is justified by the need for moral agents to provide reasons for a decision or an action to whom they are accountable.[92]

Ultimately, accountability provides a mechanism for the operationalization and allocation of responsibility in certain contexts:[93] such mechanisms applied to AI development and deployment seek to facilitate the recording, reporting, evaluation, and sanctioning of decisions and activities,[94] which contributes to building trust in the

[86] See, e.g., Miriam C Buiten, 'Towards Intelligent Regulation of Artificial Intelligence' (2019) 10 European Journal of Risk Regulation 41 <https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/towards-intelligent-regulation-of-artificial-intelligence/AF1AD1940B70DB88D2B24202EE933F1B> accessed 14 April 2023.

[87] Arguably, accountability provides the necessary framework for the requirements as set in the European Commission's White Paper (n 57).

[88] Rebecca Williams and others, 'From Transparency to Accountability of Intelligent Systems: Moving beyond Aspirations' (2022) 4 Data & Policy e7 <https://www.cambridge.org/core/journals/data-and-policy/article/from-transparency-to-accountability-of-intelligent-systems-moving-beyond-aspirations/E412FF94EC2A293985D414D80415F4AA> accessed 24 November 2022.

[89] Deven R Desai and Joshua A Kroll, 'Trust But Verify: A Guide to Algorithms and the Law' (2017) 31, p. 7-11.

[90] Paul de Hert and Guillermo Lazcoz, 'When GDPR-Principles Blind Each Other: Accountability, Not Transparency, at the Heart of Algorithmic Governance' (2022) 8 European Data Protection Law Review (EDPL) 31 <https://heinonline.org/HOL/P?h=hein.journals/edpl8&i=37> accessed 26 June 2023. See also Joshua A Kroll and others, 'Accountable Algorithms' 165 University of Pennsylvania Law Review 74, p. 658.

[91] Desai and Kroll (n 89).

[92] Mark Coeckelbergh, 'Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability' (2020) 26 Science and Engineering Ethics, p. 2052.

[93] McGrath and Whitty (n 71).

[94] Generally on how accountability may be used for governance of AI, see Theodore M Lechterman, 'The Concept of Accountability in AI Ethics and Governance' in Justin B Bullock and others (eds), *The Oxford Handbook of AI Governance* (1st edn, Oxford University Press 2022) <https://academic.oup.com/edited-volume/41989/chapter/386768252> accessed 27 February 2023.

ISSN 2281-5147

technology[95] by imposing a dimension of answerability or blameworthiness onto its developers.[96] Coherently, the NIST's recently adopted glossary for Trustworthy AI speaks the same language, linking accountability to the allocation of responsibility, and more specifically for AI governance, the obligation of an individual or an organization to account for its activities and to disclose the results in a transparent manner.[97]

Yet, the principle of accountability was, if not completely disregarded, somehow underestimated in its potential.[98] Although accountability was indeed listed among the key requirements in the Ethics Guidelines, it was presented under the (perhaps wrong) assumption that it "*complements the* [other] *requirements, and is closely linked to the principle of fairness*".[99] The following (not so wrong) assumption that accountability entails that "*mechanisms be put in place to ensure responsibility and accountability for AI systems and their outcomes, both before and after their development, deployment and use*"[100] is precisely the reason why the Ethics Guidelines missed their chance to provide tangible regulatory guidance.[101] If only accountability were regarded as a principle sharing both an 'ethical' nature *and* that of an enforcement mechanism,[102] it would have provided a precious tool for the practical implementation of the Ethics Guidelines.

---

[95] 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' (2023) Regulation & Governance <https://onlinelibrary.wiley.com/doi/abs/10.1111/rego.12512> accessed 13 February 2023.
[96] Nissenbaum (n 80).
[97] Daniel Atherton, Reva Schwartz, Peter C. Fontana, Patrick Hall (2023) The Language of Trustworthy AI: An In-Depth Glossary of Terms. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Artificial Intelligence AI 100-3. https://doi.org/10.6028/NIST.AI.100-3. <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-3.pdf> accessed 12 April 2023.
[98] The ALTAI provides a definition of accountability that is somewhat unsatisfactory: "[The] term refers to the idea that one is responsible for their action – and as a corollary their consequences – and must be able to explain their aims, motivations, and reasons. […] But accountability might also express an ethical standard, and fall short of legal consequences." The misconstruction of accountability in Artificial Intelligence is a common issue in many others so-called ethics guidelines. See, for instance, Rees and Müller (n 46), p. 6.
[99] European Commission, Directorate General for Communications Networks, and High-Level Expert Group on Artificial Intelligence (n 15), p. 19.
[100] Ibidem.
[101] Williams and others, (n 88), p. 11, "*Accountability mechanisms, on the other hand, "are procedures and tools—often technical tools, including software, but also organizational and/or legal procedures and other mechanisms—by which accountability practices are supported and implemented."*
[102] Mark Bovens, 'Two Concepts of Accountability: Accountability as a Virtue and as a Mechanism' (2010) 33 West European Politics 946 <https://doi.org/10.1080/01402382.2010.486119> accessed 2 February 2023.

ISSN 2281-5147

4. The legacy of the Ethics Guidelines in the proposed AI Act

Against this background, it is possible to draw some brief, tentative conclusions about the role that the Ethics Guidelines played in the proposed Artificial Intelligence Act (hereinafter referred to as 'AIA').[103]

It constitutes on a global scale the first attempt to regulate AI systems. Adopting a risk-based approach, the AIA identifies four categories of AI practices based on their level of social risk-acceptability, ranging from unacceptable, i.e. prohibited practices, to those that pose minimal risk. Most of the provisions in the AIA concern the 'high-risk' AI systems, for which providers are required to put a risk management system into place[104]. Although no definition of the risk management system is provided in the AIA[105], it can be entailed as a 'system' consisting of policies, procedures and instructions,[106] which shall also be approved by the responsible decision maker at the organizational level.[107] Its role is to ensure that risks are identified and adequately addressed by AI providers, despite the lack of harmonized standards or technical specifications, by adopting safeguards reducing the risks to an acceptable level.[108]

Although many considerations could be made around this piece of regulation and its foundational basis under the New Legislative Framework (NLF),[109] for the scope and

[103] European Commission, Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts COM/2021/206 Final, CELEX number: 52021PC0206, April 21, 2021.
[104] Art. 9 of the AIA.
[105] According to clause 3.2 of "ISO 31000:2018 Risk Management – Guidelines" https://www.iso.org/standard/65694.html, "risk management" can be defined as the "coordinated activities to direct and control an organisation with regard to risk".
[106] Jonas Schuett, 'Risk Management in the Artificial Intelligence Act' (2023) European Journal of Risk Regulation 1 <https://www.cambridge.org/core/product/identifier/S1867299X23000016/type/journal_article> accessed 20 February 2023; See also Comandé (n 3).
[107] In the so-called Three Lines Defense Model in the context of AI, see Jonas Schuett, 'Three Lines of Defense against Risks from AI' (arXiv, 16 December 2022) <http://arxiv.org/abs/2212.08364> accessed 13 April 2023.
[108] Ibid., p. 13.
[109] Key features of the NLF are: essential requirements the product needs to comply with to be put on the market; reliance on harmonized standards laid down by external standardization organizations to establish presumption of conformity; CE marking to certify compliance; specific obligations for the different actors along the distribution chain; conformity assessment procedure; market surveillance. See Council Resolution of 7.5.1985 on a new approach to technical harmonization and standards (OJ C 136). Regulation (EC) No 765/2008 setting out the requirements for accreditation and market surveillance relating to the marketing of products (OJ L 218). Decision No 768/2008/EC on a common framework for the marketing of products (OJ L 218).

ISSN 2281-5147

purpose of this contribution, the analysis will be narrowed down to how the seemingly 'ethical' requirements are 'translated' into legal ones for *high-risk* (emphasis added) AI systems to be complied with, reflecting the underlying need for accountability. Without claiming to be exhaustive and without delving into details, the following paragraphs briefly illustrate how most of the key requirements in the Ethics Guidelines are implemented in the legal framework, either in the AIA or in other EU legislations.

First and foremost, Article 8 of the AIA establishes that compliance with the set requirements is mandatory for high-risk systems, considering the intended purpose and the risk management system of the high-risk AI system. This means that compliance is subject to the principle of proportionality, coherent with the EU's policy statement.[110]

For the sake of clarity, the abovementioned ALTAI checklist is further analyzed, with a view to establishing the link with the requirements set under the proposed AIA.

Requirement nr. 1 related to human agency and oversight, aimed at ensuring that individuals maintain their autonomy in interacting with an AI system and a certain degree of control in the algorithmic decision-making process. The same feature is required by Article 14 of the AIA, i.e. human oversight, consisting of the adoption of appropriate human-machine interface tools, ensuring that the system is effectively overseen by natural persons during the period in which the AI system is in use.[111] The requirement of human agency and oversight can be envisioned in the transparency obligations set in Article 52 of the AIA to allow users to make informed choices for other AI systems that are not classified as 'high-risk' but are characterized by the fact that they (i) interact with humans, (ii) are used to detect emotions or determine association with (social) categories based on biometric data, or (iii) generate or manipulate content ('deep fakes').

[110] The goal is to adequately address the risks posed by AI systems in order to protect citizens, without stifling innovation, hence the proportionality principle: the higher the risks, the stricter the requirements. See, e.g., Tobias Mahler, 'Between Risk Management and Proportionality: The Risk-Based Approach in the EU's Artificial Intelligence Act Proposal' (2022) The Swedish Law and Informatics Research Institute 247 <https://www.lawpub.se/artikel/10.53292/208f5901.38a67238> accessed 13 April 2023.

[111] It has been pointed out that in data protection law, human oversight typically relates to human dignity, whereas the AIA human oversight instead relates to minimizing risks to health, safety and fundamental rights. See Frederik J Zuiderveen Borgesius, 'Strengthening Legal Protection against Discrimination by Algorithms and Artificial Intelligence' (2020) 24 The International Journal of Human Rights 1572.

ISSN 2281-5147

Requirement nr. 2 concerns technical robustness and safety, which addresses four main issues: security; safety; accuracy and reliability, fallback plans and reproducibility. Again, the same principles are 'translated' in Article 15 of the AIA in terms of accuracy, robustness and cybersecurity requirements.

Requirement nr. 3 deals with privacy and data governance, aimed at ensuring the quality and integrity of the data used, its relevance with respect to the intended purpose, and privacy protection. Needless to point out that compliance with the General Data Protection Regulation is the first step to ensure privacy and data governance. On the AIA side, the exact same requirements are mentioned in Article 10, requiring that training, validation and testing data are subject to appropriate data governance and management practices, mentioning also the need for appropriate safeguards for special categories of personal data, including technical limitations on the re-use and use of state-of-the-art security and privacy-preserving measures, such as pseudonymization or encryption.

It further specifies that datasets shall be relevant, representative, free of error[112] and complete, and with the appropriate statistical properties. These latter specifications clearly refer to requirement nr. 5 of the ALTAI, related to diversity, non-discrimination and fairness, addressing the risks of unintended prejudice and discrimination due to bias, incompleteness, and bad (data) governance models.

As far as transparency is concerned, requirement nr. 4 of the ALTAI, encompassing the need for traceability, explainability, and open communication about the limitations of the AI system is 'translated' into Articles 12 and 13 of the AIA: the former imposes record-keeping obligations for traceability purposes, the latter imposes transparency obligations to enable users to interpret output and information duties, such as characteristics, capabilities, and limitations of performance, along with appropriate instructions of use.

[112] It is true that datasets should ideally be free or error, but it would impose an unreasonably demanding standard of care onto AI providers. Recent proposals to amend Article 10 of the AIA would add a 'reasonableness filter', by stating that datasets shall be free of error "as far as this can be reasonably expected and is feasible from a technical and economical point of view". See for instance the proposed amendments by the Committee on the Internal Market and Consumer Protection Committee on Civil Liberties, Justice and Home Affairs, available here: https://artificialintelligenceact.eu/wp-content/uploads/2022/06/AIA-IMCO-LIBE-Report-All-Amendments-14-June.pdf accessed 19 April 2023.

ISSN 2281-5147

One aspect from the ALTAI checklist seems to be disregarded in the AIA, namely the societal and environmental well-being, but it is not disregarded by the legal framework altogether. In fact, the impact of AI on society and the environment, in general, is broadly addressed by the fundamental rights doctrine[113], as well as by corporate social accountability schemes[114].

Finally, the ALTAI lists accountability as requirement nr. 7, showing its close relation to risk management, identification, and mitigation in a transparent way that can be explained to and audited[115] by third parties. Coherently, in a risk-based regulation such as the AIA, it is clear thus that accountability plays a central role in the risk management framework, while at the same time providing the tools to account for trade-offs and tensions between principles and requirements laid down by the Ethics Guidelines. The rationale behind these requirements is to engage AI providers not only in the risk management process but also in active knowledge creation to prove compliance through a detailed quality management system[116] and mandatory requirements of technical documentation and record keeping[117] throughout the entire lifecycle of the product.[118] In essence, the AIA imposes a general duty onto AI providers to give explanations and justifications attesting the system's accuracy, robustness, security, transparency, and appropriate human oversight measures.

It is worth briefly mentioning that such a risk-based governance model is already familiar to the European policymaker, namely in the field of (personal) data protection, where accountability plays a central role in requiring data controllers to take responsibility for personal data processing, guaranteeing and proving compliance

[113] Art. 37 of the EU Charter of Fundamental Rights on environmental protection.
[114] See. e.g. Kudlak, Robert, Ralf Barkemeyer, Lutz Preuss, and Anna Heikkinen. *The Impact of Corporate Social Responsibility*. Milton: Taylor & Francis Group, 2022. Routledge Studies in Management, Organizations and Society. See also Weslei Gomes de Sousa and others, 'How and Where Is Artificial Intelligence in the Public Sector Going? A Literature Review and Research Agenda' (2019) 36 Government Information Quarterly 101392 <https://www.sciencedirect.com/science/article/pii/S0740624X18303113> accessed 19 April 2023.
[115] Note that the ALTAI mentions the need for auditability of AI systems to ensure independent evaluation of AI systems. Although no specific provision concerning auditing is contained in the AIA, Recital 69 calls for the elaboration of functional specifications by the European Commission and independent audit reports with regard to the EU database for high-risk AI systems.
[116] Art. 17 of the AIA.
[117] Art. 11 and 12 of the AIA.
[118] de Hert and Lazcoz (n 90), p. 38.

ISSN 2281-5147

with the principles set in article 5 of the General Data Protection Regulation[119] (hereinafter referred to as the 'GDPR'). Here again, what may seem as the last *and* least principle among those governing data protection is actually – again – an overarching principle linked to the others listed in article 5 of the GDPR, namely lawfulness, fairness and transparency, purpose limitation, data minimization, accuracy, storage limitation, integrity, and confidentiality, allowing for a process-oriented demonstration of compliance.[120]

In conclusion, although the AIA has not entered into force yet, providers of *all* (emphasis added) AI systems should not turn to AI ethics or 'ethics' broadly speaking, but to the principle of accountability to develop their AI systems. To provide further corroboration, consider how, pending the AIA approval in the trialogue, the European Parliament proposed an amendment to the original legislative text as to include *ex novo* Article 4 a, indexed as "*General principles applicable to all AI systems",* which reports at paragraph 2 from a) to f) the seemingly ethical guidelines for Trustworthy AI.[121] However, such attempt to revive the HLEG's Ethics Guidelines shall not be interpreted as an exemption from adequate risk-management for non-high-risk AI systems. On the contrary, the expected 'best effort' shall be interpreted as an accountability measure to ensure the adoption of the highest possible standards for risk management, considering the state-of-the-art technological solutions and the requirements laid down in the AIA, regardless of the formal classification of risk.

As such, the assessment of compliance with the requirements for Trustworthy AI is exempt from any evaluation of an 'ethical' nature but is to be carried out from a perspective of risk regulation.[122] A comprehensive risk assessment for AI systems would consist in identifying and estimating the likelihood of risks occurring, analyzing

[119] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).
[120] de Hert and Lazcoz (n 90), p. 38.
[121] Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)), available at: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html, accessed 25 November 2023.
[122] Kaminski (n 78).

ISSN 2281-5147

their potential impact, carrying out trade-offs, followed by implementing risk management actions to mitigate their effects on society.[123]

5. Accountability: a common language for AI regulation

AI regulation is essentially a matter of social risk acceptability:[124] to determine what is seen as acceptable[125] requires making tensions and trade-offs explicit and subject to interrogation, demonstration, debate, and justification,[126] and such a regulatory action speaks the language of accountability[127], which ultimately contributes to the societal trust in the technology.[128]

To achieve this objective, the United States is also actively working towards establishing a culture of efficient risk management as part of its overall approach to AI regulation. The National Institute of Standards and Technology (NIST) has developed the AI Risk Management Framework (referred to as 'AI RMF')[129], which approaches risk management in the context of AI through the perspective of enterprise risk management. Devised based on the NIST's Cybersecurity Framework of 2014[130], the AI RMF will consist of a Core and Profiles, where companies will use the Core to identify desired outcomes and techniques, select a risk-management Profile, and operationalize risk management by adopting suitable practices based on their preferences. However, unlike the European centralized top-down and precautionary approach, the AI RMF is to be regarded as soft law, i.e. mandatory

[123] Christoph Lütge and others, 'On a Risk-Based Assessment Approach to AI Ethics Governance', IEAI White Paper (2022), p. 1.
[124] Simone Borsci and others, 'Embedding Artificial Intelligence in Society: Looking beyond the EU AI Master Plan Using the Culture Cycle' (2022) AI & SOCIETY <https://doi.org/10.1007/s00146-021-01383-x> accessed 23 January 2023.
[125] See generally Bridget Hutter, 'The Attractions of Risk-Based Regulation: Accounting for the Emergence of Risk Ideas in Regulation' (2005).
[126] See the definition provided by the ALTAI (n 98). See also Francesco Gualdi and Antonio Cordella, 'Artificial Intelligence and Decision-Making: The Question of Accountability' (2021) <http://hdl.handle.net/10125/70894> accessed 12 August 2022.
[127] See, e.g., Kirsten Martin, 'Ethical Implications and Accountability of Algorithms' (2019) 160 Journal of Business Ethics 835 <https://doi.org/10.1007/s10551-018-3921-3> accessed 12 August 2022.
[128] Laux, Wachter and Mittelstadt (n 95).
[129] U.S. Department of Commerce, National Institute of Standards and Technology (NIST), Artificial Intelligence Risk Management Framework (AI RMF 1.0), 2023, available at: https://www.nist.gov/itl/ai-risk-management-framework accessed 13 July 2023.
[130] See the 2018 version 1.1 NIST Framework for Improving Critical Infrastructure Cybersecurity, available at: https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf accessed 13 July 2023.

ISSN 2281-5147

guidance on risk mitigation strategies.[131] It nonetheless contemplates AI risk categorization and the adoption of technical and organizational measures for risk mitigation throughout the lifecycle of an AI system, and thus highlights the importance for organizations to establish and maintain appropriate accountability mechanisms, defining roles, responsibilities, and incentive structure for an effective risk-management framework.[132]

The aforementioned 'accountability language' is also spoken by the US proposed Algorithmic Accountability Act (hereinafter 'AAA'). The self-explanatory title of the bill, first introduced in the US Congress in 2019, then revised in 2022,[133] can be seen as an indication that policymakers in the US are increasingly aware of the importance of the principle of accountability in AI regulation. More specifically, the US policymaker is concerned with regulating 'critical decision processes' involving algorithmic systems that may have a significant legal or material effect primarily on consumers: this is one major difference compared to the EU's AIA, which regulates AI systems in general.[134]

If passed into law, the Act would require covered entities,[135] i.e. large companies deploying automated decision systems or augmented critical decision processes, to carry out impact assessments of their AI systems on a range of factors, including bias, fairness, and privacy. As mentioned above, impact assessments are but one tool from the regulatory toolkit provided by the principle of accountability, plus there are various models on how to carry out an algorithmic impact assessment[136]. Nonetheless, the impact assessment model of the AAA resembles that of a collaborative

[131] Kaminski (n 78).
[132] AI RMF (n 130), p. 9.
[133] The Algorithmic Accountability Act was introduced by senators Ron Wyden, Cory Booker, and representative Yvette Clarke, with the goal of setting transparency and oversight for automated decision-making systems that affects mainly consumers. <https://www.wyden.senate.gov/news/press-releases/wyden-booker-and-clarke-introduce-algorithmic-accountability-act-of-2022-to-require-new-transparency-and-accountability-for-automated-decision-systems> accessed 5 May 2023. An important feature, which stresses the consumer protection purpose of the Act, is contained in Section 9, where it is stated that a violation of the Act or a regulation promulgated thereunder shall be treated as a violation of a rule defining an unfair or deceptive act or practice under section 18(a)(1)(B) of the Federal Trade 15 Commission Act (15 U.S.C. 57a(a)(1)(B)).
[134] Jakob Mökander and others, 'The US Algorithmic Accountability Act of 2022 vs. The EU Artificial Intelligence Act: What Can They Learn from Each Other?' (2022) Minds and Machines <https://doi.org/10.1007/s11023-022-09612-y> accessed 21 November 2022.
[135] AAA, Sec. (2)(7).
[136] For a thorough study on capabilities and limitations of various algorithmic impact assessment models see Andrew D Selbst, 'An Institutional View of Algorithmic Impact Assessments', 35 HARV. J.L. & TECH. 117 (2021).

ISSN 2281-5147

governance approach, similar to the GDPR, consisting of both a top-down action under the Federal Trade Commission's ('FTC') regulatory oversight and internal organizational compliance culture.[137] The requirements for such impact assessment are contained in eleven paragraphs,[138] the thorough analysis of which would go beyond the scope of this article. Nonetheless, it is worth noting that most requirements deal with specifications around its performance by the covered entities, taking into specific account the attempt to eliminate and mitigate any impact that is likely to cause a material negative impact with legal or similarly significant effects on consumers,[139] and the maintenance of relevant documentation. Moreover, the AAA requires covered entities to perform testing related to potential discriminatory impact, showing how the policymaker is concerned that the delegation of critical decision processes to private companies is oriented towards the public good.[140]

Although the impact assessment is not overseen by regulators and thus it is not subject to public scrutiny, another accountability measure is implemented: covered entities are required to prepare and submit a summary report of the impact assessment to the FTC, which in turn is required to create a public repository containing substantial information from such summary reports.[141]

However, the AAA does not adopt a precautionary approach to AI risk categorization, prohibiting or strictly regulating conditions for high-risk AI applications, and it relies on self-assessments to be conducted by the providers themselves, as such substantially differing from the EU's AIA, which heavily relies on harmonized standards.[142]

Overall, the same accountability vocabulary in terms of standardization, privacy and non-discrimination, information security measure, testing and evaluation of the

[137] Kaminski (n 78), p. 63.
[138] AAA, Sec. 3.
[139] AAA, Sec. 3(b)(1)(H).
[140] Furkan Gursoy, Ryan Kennedy and Ioannis Kakadiaris, 'A Critical Assessment of the Algorithmic Accountability Act of 2022' (3 March 2022) <https://papers.ssrn.com/abstract=4193199> accessed 13 July 2023.
[141] AAA, Sec. 6(b).
[142] Gursoy, Kennedy and Kakadiaris (n 140), p. 4-5.

ISSN 2281-5147

model's performance, documentation, and data quality is used to describe the covered entities' requirements in Section 4 of the AAA.[143]

On a more general note, it is worth noting that the OECD, besides developing its own AI principles for Trustworthy AI[144], has also recently proposed an accountability framework for AI risk governance and management.[145] It first and foremost states the need to clearly define roles and responsibilities of AI developers towards the deployers and all the possible downstream applications. The OECD accountability framework extends to all actors in the AI ecosystem, who should manage risks according to their roles, context, and following state-of-the-art practices.[146] This entails designing, implementing, and overseeing processes that encompass documenting AI system decisions, enabling audits, and providing adequate responses to risks and redress mechanisms.[147]

Furthermore, it outlines the necessity to manage risks throughout the entire AI systems' lifecycle, which includes planning, designing, data collection and processing, model building and validation, deployment, operation, and monitoring. This can be performed through a four-step process, briefly reported as follows: i) Define: establishing the scope, context, actors, and criteria for evaluating AI system risks; ii) Assess: identifying, evaluating, and measuring risks to ensure the AI system functions as intended and remains trustworthy; iii) Treat: implementing appropriate techniques to prevent, mitigate, or cease identified problems; iv) Govern: ensuring continual monitoring, reviewing, documenting, communicating, and consulting on AI risk management actions and outcomes. Finally, the report emphasizes the importance of

---

[143] See generally AAA, Sec. 4. More particularly, at Sec. 4(a)(11), according to which covered entities are required to "[i]*dentify any capabilities, tools, standards, datasets, security protocols, improvements to stakeholder engagement, or other resources that may be necessary or beneficial to improving the automated decision system, augmented critical decision process, or the impact assessment of such system or process, in areas such as— (A) performance, including accuracy, robustness, and reliability; (B) fairness, including bias and non-discrimination; (C) transparency, explainability, contestability, and opportunity for recourse; (D) privacy and security; (E) personal and public safety; (F) efficiency and timeliness; (G) cost; or (H) any other area determined appropriate by the Commission."*

[144] OECD (n 27).

[145] OECD, 'Advancing Accountability in AI: Governing and Managing Risks throughout the Lifecycle for Trustworthy AI', vol 349 (2023) OECD Digital Economy Papers 349 <https://www.oecd-ilibrary.org/science-and-technology/advancing-accountability-in-ai_2448f04b-en> accessed 16 August 2023. Interestingly, at the very beginning the report answers the questions of what constitutes Trustworthy AI, immediately followed by the concept of accountability in AI.

[146] Ibid., p. 17.

[147] Ibid, pp. 22-23.

ISSN 2281-5147

instilling a culture of risk management throughout organizations and along the entire AI value chain.

As such, it is safe to consider that the OECD has to some extent provided a comprehensive taxonomy of accountability tools, which can be safely adopted by policymakers.

5.1. Accounting for possible mistranslations: the case of foundation models

Contextually to the proposed Article 4 a), following the widespread adoption of ChatGPT, released in early November 2022[148], along with rising concerns over the additional risks posed by generative AI,[149] the European Parliament has proposed amendments to specifically address foundation models,[150] in an attempt to counter the premature obsolescence of the AIA. Therefore, numerous references to foundation models were included, notably from Recital 60 e) to 60 h), along with Article 28 b) that imposes specific obligations to developers of foundation models. The most relevant requirement for the analysis at hand is the technical documentation related to the capabilities and limitations, the development, the testing and the validation of the foundation model that the provider must make available to the downstream developers. This is to be regarded as a risk mitigation strategy, accounting for the complex value chain of an AI system.

Nonetheless, the complexity of an AI value chain is not new, and the recent uptake of foundation model has only added an additional layer of complexity to a problem that was already well-known. The policymaker's - and more specifically the European Parliament's - introduction of additional obligations onto providers of foundation

[148] OpenAI, Introducting ChatGPT, available at: https://openai.com/blog/chatgpt accessed 25 November 2023.

[149] Consider, for example, how on 30 March 2023 the Italian Data Protection Authority temporarily banned ChatGPT for Italian users over privacy concerns. See Garante per la Protezione dei Dati Personali, Provvedimento n. 9870832 del 30 marzo 2023, available at: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9870832 accessed 25 November 2023.

[150] Foundation models refers to a type of AI system that is developed using extensive amount of data, designed to produce a wide variety of outputs, and capable of being customized for a wide range of downstream and more specific tasks. Large Language Models, such as GPT by OpenAI that powers ChatGPT, are a type of foundation models capable of generating text.

ISSN 2281-5147

models is but an attempt to solve the so-called 'many hands problem',[151] which refers broadly to the challenge of attributing responsibility in complex organizational settings where many individuals contribute to decisions and outcomes, making it difficult to pinpoint who is accountable. The additional obligations set in Article 28 b) are meant to solve the 'many hands problem' by establishing a clear framework of roles and responsibilities along the actors involved in the AI system's value chain as to include the developers of the foundation model, who would stand at the very beginning of the development pipeline.

However, this apparently noble attempt may end up causing a backlash effect by exposing the EU regulatory approach to AI to the 'pacing problem',[152] while at the same time undermining the normative force of the very same principles of Trustworthiness it aims to enforce. On the one hand, specific provisions covering foundation models may be a suitable regulatory response to the societal concerns about generative AI, but they might only offer a temporary relief for a permanent problem. Foundation models are just the frontier technology; but if the AIA is expected to be the frontier regulation, it must keep the pace of innovation beyond "the law of the horse",[153] or the law of foundation models in Article 28 b) of the AIA. On the other hand, the return to general principles of Trustworthy AI envisioned in Article 4 a) may indeed constitute a permanent solution to yet emerging problems, but only if interpreted through the lenses of accountability, as argued *supra*. Failing in doing so would not only lead to redundant regulatory provisions but would also practically diminish the normative capability of such principles for a future-proof and resilient regulation of technology.[154]

The central conflict lies in the yet-to-be-achieved balance between principle-based regulation and rule-based regulation for AI risk management.[155] While the latter is

[151] Nissenbaum (n 80).
[152] See generally Gary E Marchant, 'The Growing Gap Between Emerging Technologies and the Law' in Gary E Marchant, Braden R Allenby and Joseph R Herkert (eds), *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight: The Pacing Problem* (Springer Netherlands 2011) <https://doi.org/10.1007/978-94-007-1356-7_2> accessed 26 November 2023.
[153] Frank H Easterbrook, 'Cyberspace and the Law of the Horse' (1996) The University of Chicago Legal Forum.
[154] Gary E Marchant and Yvonne A Stevens, 'Resilience: A New Tool in the Risk Governance Toolbox for Emerging Technologies' (2017) 51 U.C. DAVIS L. REV. 2.
[155] See generally Ruth B Carter and Gary E Marchant, 'Principles-Based Regulation and Emerging Technology' in Gary E Marchant, Braden R Allenby and Joseph R Herkert (eds), *The Growing Gap Between Emerging Technologies*

ISSN 2281-5147

criticized for stifling innovation by allegedly imposing excessive compliance costs and burdens, which has led to France, Germany, and Italy advocating for self-regulation through codes of conduct for providers of foundation models,[156] the former requires clearer interpretation to guarantee enforcement. This situation can be seen as a case of evident mistranslation, revealing the potential shortcomings of an accountability-based regulation. If not properly understood by the policymakers themselves, it could lead the AI regulation towards one of the extremes: either inserting redundant provisions for any last-minute piece of technology or reverting to the same 'ethics washing' approach that was thought to have been lost in translation once and for all.

6. Conclusive remarks

In conclusion, the previous analysis illustrated the evolution of AI regulation, with a particular focus on the European Union's approach to AI ethics and the transition to the principle of accountability. The EU's AI Ethics Guidelines, developed by the HLEG, have served as a foundation for ethical and trustworthy AI development and deployment within the European Union. However, the paper has highlighted the shortcomings of these guidelines when applied as regulatory tools, leading to possible phenomena of 'ethics washing'. Therefore, it exposes the need for a more robust and legally enforceable approach.

The discussion of the proposed Artificial Intelligence Act has illustrated how the seemingly 'ethical' principles from the Ethics Guidelines are translated into requirements for high-risk AI systems under the AIA, reflecting the underlying need for accountability. The principle of accountability emerged as a common language for AI regulation, contributing to societal trust in technology and serving as a foundation for regulatory frameworks around the world. On the same wavelength, in the context of the US-proposed Algorithmic Accountability Act and the AI RMF, we have demonstrated that the same accountability vocabulary is used to describe the

and Legal-Ethical Oversight: The Pacing Problem (Springer Netherlands 2011) <https://doi.org/10.1007/978-94-007-1356-7_10> accessed 26 November 2023.

[156] Andreas Rinke, 'Exclusive: Germany, France and Italy reach agreement on future AI regulation' (Reuters, 20 November 2023), available at: https://www.reuters.com/technology/germany-france-italy-reach-agreement-future-ai-regulation-2023-11-18/ accessed 25 November 2023.

ISSN 2281-5147

requirements for covered entities, emphasizing the importance of the principle of accountability in AI regulation, both in the US and the EU.

Ultimately, we argued that *"Trustworthy AI depends upon accountability".*[157] AI providers should focus on the principle of accountability to develop their AI systems, accounting for the highest possible standards for risk mitigation measures, in line with what is already established in the field of personal data protection. The Ethics Guidelines may have provided a common global consensus on AI governance, but due to its lack of enforcement mechanisms, the strictly 'ethical' approach got lost in translation from AI governance to AI regulation. Arguably it is the principle of accountability that provides a common language, both in the EU and the US, which translates the need for transparency, fairness, explainability, and human oversight into practical technical requirements.

Where the Ethics Guidelines fell short, accountability bridges the gap between the abstract 'ethical' principles for trustworthy, *rectius* responsible, AI and the regulatory framework by providing the necessary implementation mechanism that guarantees compliance, oversight, monitoring, verification, and documentation of the AI system's development.

[157] Directly quoted from the AI RMF (n 130), p. 15.

ISSN 2281-5147