A Voluntary Certification Mechanism for Public Interest AI

Exploring the Design and Specifications of Trustworthy Global Institutions to Govern AI

RESEARCH PAPER

SEPTEMBER 2024

Forum on Information & Democracy

EXECUTIVE SUMMARY

The pace of artificial intelligence (AI) development and deployment is unprecedented, greatly transforming societies and democracies. The mission of the Forum on Information and Democracy is to recommend policy solutions to ensure democratic safeguards in the global information and communication space. This paper explores trustworthy global institutions to govern AI, notably in investigating the creation of a voluntary certification mechanism for public interest AI, as suggested in the Forum's report *AI as a Public Good: Ensuring Democratic Control of AI in the Information Space.*[1]

The paper argues that a voluntary certification mechanism, built on the success of the Fairtrade mechanism, could provide the necessary incentives to encourage the development and deployment of AI systems built in the public interest. Public interest AI are AI systems directly accessible to users, which enhance the goals all the members of the public share. The paper explores the potential contributions of such a mechanism, which include tackling the information asymmetry between users and AI providers, offering a solution when regulation is insufficient or absent, and responding to a fast-evolving transnational challenge. The mechanism, suggested to be led by civil society, would also create a system in which the definition of what constitutes a public interest AI system is not in the hands of private actors but the users.

This mechanism would focus on AI providers and their systems, as they are responsible for the deployment of the system accessible to the public. This focus notwithstanding, the creation of a complementary voluntary certification system for AI foundation models could be envisioned.

The paper then moves on to explore the certification criteria that should be assessed to identify a public interest AI system. It argues that the functioning and values of the AI system provider are an important proxy to identify the public interest character of the system. These include notably the provider's official commitment to human rights, democratic and transparent governance, employment and working conditions, and environmental sustainability. Other criteria such as anti-corruption policies could be further explored. In addition to these criteria, criteria regarding the system should be taken into account. These include the public interest objective of the system, the respect for fairness in data curation and use, privacy, and non-discrimination. The paper argues that the transparency of the system as well as the setup of inclusive and participatory processes to develop the AI system are of particular relevance, as they enable external scrutiny and collective decision-making with regard to acceptable and unacceptable risks, bias, and security concerns.

Finally, the paper explores the potential governance of the certification mechanism, which is equally important as the assessment criteria, notably with regards to legitimacy, integrity, and uptake of the mechanism. Taking inspiration from the ISEAL Credibility Principles, it argues that to set the standards, the mechanism should build on internationally recognised AI principles and guidelines and should be managed by an independent and non-profit organisation with 50% Global South representation. It discusses the inclusion of industry and government actors, which on the one hand create risks of capture, but which on the other are needed to avoid backlash and prepare for the creation of alternative labels. The paper comes to the conclusion that a symbolic inclusion of these actors could be envisioned.

1 Published in February 2024 by the Forum on Information and Democracy, the policy framework *AI as a Public Good: Ensuring Democratic Control of AI in the Information Space* was co-chaired by Laura Schertel Mendes, Brazilian lawyer, and Jonathan Stray, computer scientist at the University of Berkeley.

The paper further suggests a two-step certification process which includes a self-assessment component with a learning objective, as well as certification through an independent body, accredited in line with ISO/IEC 17065 to create trust. It highlights the importance of monitoring and enforcement processes to avoid fraud and ensure integrity of the label, as well as the need for a sustainable and independent financing model which could rely on a fee to be paid by the certified AI system providers.

In order to foster uptake of the certification mechanism, awareness-raising campaigns to create consumer demand, public procurement and trade policies, as well as legal and political recognition of the label are recommended. The paper explores in particular how the mechanism could be aligned with the forthcoming OECD *Due Diligence Guidance for Responsible AI* and UNESCO's *Ethical Impact Assessment*. Finally, it argues that continuous monitoring and learning are needed to adapt the mechanism to new developments in AI.

The paper thereby provides the basis for more in-depth discussions among relevant actors about the creation of a voluntary certification mechanism for public interest AI as one solution to create trustworthy institutions to govern AI.

TABLE OF CONTENT

ACKNOWLEDGMENTS

The Forum on Information and Democracy wishes to express its gratitude to all actors who made this research paper possible.

We would specifically like to thank Marta Cantero Gamito, Professor of IT Law, University of Tartu and research fellow, Florence School of Transnational Governance at the European University Institute. Marta has supported the drafting process of this paper and first suggested the idea of a certification mechanism for AI inspired by the Fairtrade mechanism, in the framework of the work on the Forum's report *AI as a Public Good: Ensuring Democratic Control of AI in the Information Space*, published in February 2024.

The Forum would also like to express its gratitude to all the members of its Working Group on AI and its implication for the information space, who provided their insights for the report, and to the rapporteurs who drafted it. The report is the basis for this research paper.

The Forum further would like to express its gratitude to all actors who provided useful insights for this paper in sharing their expertise and information about their own initiatives namely,

- **Rashad Abelson and Elisa Narminio,** OECD
- **Rachel Adams,** Global Index on Responsible AI
- **Vincent Berthier and Arthur Grimonpont,** Reporters without Borders (RSF)
- **Blaise Desbordes,** Max Havelaar – Fair Trade France
- **Audrey Herblin,** Mistral
- **Syed Nazakat,** DataLeads
- **Dan-Flaviu Nechita,** former Head of Cabinet for Dragoș Tudorache, Member of European Parliament
- **Claire Jenifer Pershan,** Mozilla Foundation
- **Alejandro Pisanty,** Professor at UNAM, Mexico, former member of the Board of Directors at ICANN
- **Ed Newton-Rex,** Fairly Trained
- **Gabriela Ramos,** UNESCO
- **Benjamin Sabbah and Chloe Fiodiere,** Journalism Trust Initiative
- **Aurélien Sautière,** Forest Stewardship Council
- **Andrea Schlüter,** TÜV AI.Lab
- **Nina Santos,** Aláfia Lab
- **Stefanie Vehring,** standardisation expert

This research paper was produced by the Forum on Information and Democracy, chaired by Thibaut Bruttin. The paper was elaborated by Katharina Zuegel, Policy Manager, under the strategic direction of Camille Grenier, Executive Director. Celia Rochepitard supported the research and drafting process.

The Forum wishes to thank the Future of Life Institute for its financial support which enabled this project in the context of its research *Exploring the Design and Specifications of Trustworthy Global Institutions to Govern AI*.

ABOUT THE FORUM ON INFORMATION AND DEMOCRACY

Established in 2019, the Forum on Information and Democracy, is the implementing organisation of the Intergovernmental Partnership for Information and Democracy, endorsed to date by 54 States.

This Partnership sets high-level principles to promote and implement democratic safeguards in the global information and communication space. It strives to enhance intergovernmental cooperation on information integrity, ranging from press freedom and the sustainability of journalism to platforms governance and digital rights.

As the Secretariat of the Partnership, the Forum is mandated to evaluate the means, norms and architecture of the information and communication space and facilitate the emergence of regulatory and self-regulatory responses.

It has published hundreds of policy recommendations on how to end infodemics (2020), how to launch a new deal for journalism (2021), how to establish accountability regimes for social networks and their users (2022), how to ensure pluralism of algorithms (2023), and AI as a public good (2024).

The Forum also hosts the Observatory on Information and Democracy, an equivalent of what the IPCC represents for climate change, which aims to evaluate the means, norms and architecture of the information and communication space and their impacts on democracy.

INTRODUCTION

The pace of artificial intelligence development and deployment is unprecedented, greatly transforming societies and democracies. Companies are in a race to build Artificial General Intelligence (AGI), which will outperform human capabilities. Yet, these transitions are mainly led by private companies acting in their own interest which put ethical considerations at most as a secondary priority. International organisations, governments and civil society are thus calling for trustworthy AI and global governance mechanisms to oversee development and deployment of AI systems. Specifically, as part of its "Realising Aspirational Futures" programme, the Future of Life Institute launched a call for research projects that investigate potential global mechanisms or institutions to govern advanced AI in the near future.

This paper, published by the Forum on Information and Democracy with the support of the Future of Life Institute, is a contribution to the reflections on global governance mechanisms. It is inscribed in the mandate of the Forum, which aims to promote democratic safeguards in the global information and communication space. The Forum is the implementation body of the International Partnership for Information and Democracy, endorsed by 54 countries worldwide.

This paper is a continuation of the Forum's policy framework, *AI as a Public Good: Ensuring Democratic Control of AI in the Information Space,* published in February 2024. The report provides more than 200 policy recommendations to safeguard information integrity in the AI age. It notably suggests the creation of a voluntary certification mechanism for AI systems inspired by the Fairtrade mechanism as a means to encourage the development of trustworthy AI systems.

In this regard, this paper explores what a voluntary certification mechanism for AI systems directly accessible to users could look like. It suggests the creation of a public interest AI certification mechanism which puts the public interest character of AI systems at the heart of AI development and use. It explores the criteria that such a mechanism should assess to ensure that an AI system is built in and for the public interest. It also explores a potential governance system for the certification mechanism, i.e., the scheme owner. This paper builds on the premise that a focus on the public interest enables minimising risks of advanced AI systems and maximising benefits to society.

The paper further suggests aligning the voluntary certification mechanism with existing and ongoing efforts by international organisations to ensure legitimacy, sustainability, and uptake. In particular, this concerns UNESCO's *Ethical Impact Assessment,* which supports government officials involved in public procurement of AI systems. The proposed public interest AI certification mechanism could be considered as a means to certify the ethics of an AI system. It also concerns the forthcoming OECD *Due Diligence Guidance for Responsible AI,* for which the proposed public interest AI certification mechanism could become an assessment tool.

This paper is an exploratory work which does not aim to have the answers to all questions, but rather provides a basis for a more in-depth reflection on the establishment of a voluntary certification mechanism for AI systems as one of the solutions to advance trustworthy AI. Actors from civil society, government, research and international organisations and others are encouraged to further explore the idea and implement the project.

1. AN INNOVATIVE APPROACH TO GOVERNING AI SYSTEMS: SETTING-UP A VOLUNTARY CERTIFICATION MECHANISM

1.1/ WHAT IS A VOLUNTARY CERTIFICATION MECHANISM?

Most voluntary certification mechanisms fall into the category of sustainability initiatives, identified by the OECD and defined as any NGO, multistakeholder, government-backed or industry initiative, scheme or programme that sets requirements for, or assesses business practices in relation to sustainability objectives (i.e., objectives related to human rights, social or environmental impacts).[2]

In addition, the voluntary character of such a mechanism needs to be underlined. According to Depoorter and Marx, voluntary sustainability standards (VSSs) have developed markedly since the 1990s. They emerged as transnational, private, market-based regulatory instruments in which a group of actors developed and applied standards related to production practices in order to tackle the sustainability issues pervasive in global value chains.[3] A voluntary certification scheme represents first and foremost a transnational, private regulatory instrument characterised by the voluntary nature of its adherents.

Finally, this mechanism is characterised by being sanctioned by an accredited certifying body. In this respect, the International Standards Organisation (ISO) defines certification as "the provision by an independent body of written assurance that the product, service or system in question meets specific requirements".[4]

These VSSs have proliferated, with more than 300 currently in existence, such as Fairtrade or the Forest Stewardship Council® (FSC®).[5] This proliferation has been accompanied by research on its impact. While, for example, Fairtrade had an impact on awareness and consumption, it still remains to a certain extent a niche product.[6] Studies on the FSC show that the certification system has had an impact on environmental sustainability, working and living conditions and legislation.[7] However, certified forests

---

2 OECD (2022), *The role of sustainability initiatives in mandatory due diligence: Background note on Regulatory Developments concerning Due Diligence for Responsible Business Conduct*, available at : https://mneguidelines.oecd.org/the-role-of-sustainability-initiatives-in-mandatory-due-diligence-note-for-policy-makers.pdf (accessed 28 August 2024).

3 Depoorter, C. and Marx, A. (2023), Fostering compliance with voluntary sustainability standards through institutional design: An analytic framework and empirical application, *Wiley*, available at : www.evidensia.eco/resources/2278/fostering-compliance-with-voluntary-sustainability-standards-through-institutional-design-an-analytic-framework-and-empirical-application/ (accessed 28 August 2024).

4 ISO,*Conformity assessment Certification*, available at : https://www.iso.org/certification.html (accessed 28 August 2024).

5 Depoorter, C. and Marx, A. (2023), Fostering compliance with voluntary sustainability standards through institutional design: An analytic framework and empirical application, *Wiley*, available at : www.evidensia.eco/resources/2278/fostering-compliance-with-voluntary-sustainability-standards-through-institutional-design-an-analytic-framework-and-empirical-application/ (accessed 28 August 2024).

6 Vázquez, J-L., Lanero, A., García, J. A., Moraño, X (2023), *Segmentation of consumers based on awareness, attitudes and use of sustainability labels in the purchase of commonly used products*, available at: www.evidensia.eco/resources/2206/segmentation-of-consumers-based-on-awareness-attitudes-and-use-of-sustainability-labels-in-the-purchase-of-commonly-used-products/ (accessed 28 August 2024) & Redfern, A. and Snedker, P. (2002), *Creating Market Opportunities for Small Enterprises: Experiences of the Fair Trade Movement, ILO,* available at: www.ilo.org/publications/creating-market-opportunities-small-enterprises-experiences-fair-trade (accessed 28 August 2024).

7 WWF (2014), *The impact of Forest Stewardship Council Certification*, available at: https://wwfint.awsassets.panda.org/downloads/wwf_impact_forest_stewardship_council_certification_2014.pdf (accessed 28 August 2024).

represent only 5% of forests worldwide. While an in-depth assessment of the impact of such systems goes beyond the scope of this paper, it is important to note that overall, VSSs have brought ethical issues to mainstream trade and business practices and encouraged the adoption of legislation.[8]

EXAMPLES OF VOLUNTARY CERTIFICATION MECHANISMS THAT PROVIDE INSPIRATION FOR AN AI CERTIFICATION MECHANISM

BLUE ANGEL

This environmental label was developed and is owned by the German Federal Ministry on Environment in cooperation with the Federal Environmental Agency. In the framework of the Blue Angel certification, the agency developed a label evaluating the environmental impact of software products, the "Resource and Energy-Efficient Software Products". It focuses on three criteria notably transparency about the software's energy consumption, saving resources through reducing hardware requirements, and user autonomy enabling users to reduce the environmental impact.[9]

FAIRTRADE INTERNATIONAL

Fairtrade International is a non-profit, multi-stakeholder organisation that has established one of the world's most renowned sustainability labels. Resolutely focused on the fair treatment of small producers, this label requires fair remuneration, decent working conditions, respect for the environment, and the autonomy and democratic management of producer cooperatives. The label is obtained via a voluntary certification mechanism, administered by Flocert, the certifying body, which carries out investigations in order to ultimately deliver the certification.[10]

FAIRLY TRAINED

Launched in January 2024, Fairly Trained is a voluntary certification not-for-profit, which identifies generative AI companies that train their systems respecting data creators' rights through consent, instead of building upon the concept of fair use. It considers systems to be fairly trained if the data is available in the public domain, is under an open licence, or if a contractual agreement has been established. By issuing its label, Fairly Trained encourages AI companies to train their AIs with the creators' consent.[11]

FOREST STEWARDSHIP COUNCIL (FSC)

Created in 1994 following the United Nations Conference on Environment and Development, the international NGO Forest Stewardship Council (FSC) launched a voluntary certification scheme designed to protect forests. Certification is issued after an audit conducted by an accredited body. Two particular features characterise this

8 Redfern, A. and Snedker, P. (2002), *Creating Market Opportunities for Small Enterprises: Experiences of the Fair Trade Movement*, ILO, available at: www.ilo.org/publications/creating-market-opportunities-small-enterprises-experiences-fair-trade (accessed 28 August 2024).
9 Blue Angel (2020), *Resource and Energy-Efficient Software Products, Basic Award Criteria* Edition January 2020 Version 3, available at: https://produktinfo.blauer-engel.de/uploads/criteriafile/en/DE-UZ%20215-202001-en-Criteria-V3.pdf (accessed 28 August 2024).
10 Fairtrade International, available at: www.fairtrade.net (accessed 29 August 2024).
11 Fairly Trained, available at: www.fairlytrained.org/ (accessed 28 August 2024).

mechanism. The first is its complex governance, involving different local and hierarchical levels and taking into account the points of view of a variety of stakeholders (NGOs, civil society, trade unions, industrialists, and retailers). Second, it employs a multiplicity of labels, depending on the typology of products: "100%", "Recycled", and "Mixed".[12]

JOURNALISM TRUST INITIATIVE (JTI)

The Journalism Trust Initiative is an international, independent mechanism designed as an ISO standard for recognising ethical journalistic practices. Launched by Reporters Without Borders (RSF), the initiative has established 130 criteria and already includes over 1,700 media outlets in 100 countries. The certification process is voluntary, and involves two steps: a self-assessment process which also works as a pedagogical process to bring the media up to standard, followed by an external audit conducted by an accredited certification body.[13]

1.2/ POTENTIAL CONTRIBUTIONS OF A VOLUNTARY CERTIFICATION MECHANISM TO GOVERN AI SYSTEMS

The development and adoption of AI systems in the public domain has proliferated and accelerated, particularly since the release of ChatGPT. This acceleration does not only have important implications for the information and communication space and democracy, as discussed in the Forum's report,[14] but it also announces important long-term changes as actors are "pursuing artificial general intelligence which can perform as well as or better than humans at a wide range of tasks".[15]

A voluntary certification mechanism to promote and identify ethical and responsible AI systems working in the public interest could therefore reply to several challenges:

- ***An indicator for consumers to tackle information asymmetry***

As analysed by the UK's Department for Science, Innovation and Technology, a large information asymmetry exists between those developing AI systems, particularly frontier models, and those regulating and using them.[16]

This information asymmetry between sellers and buyers, as theorised by Georges Akerlof, leads to a deterioration of quality of products. Akerlof explains in his theory of "The Market for Lemons" that buyers who do not have information about a product's quality will not be able to identify a high-quality product and will judge the product by the price, leading to a deterioration of the quality of products.[17] Even though this theory was developed for the used car market, it can also apply to the AI market. The lack of information about AI systems, their societal impact, and trustworthiness inhibits users, including consumers, public administration, and others from choosing the best product. To counter this

12 Forest Stewardship Council, available at: https://fsc.org/en (accessed 29 August 2024).
13 JTI, Journalism Trust Initiative, available at: www.journalismtrustinitiative.org (accessed 28 August 2024).
14 Forum on Information and Democracy (2024), *AI as a Public Good: Ensuring Democratic Control of AI in the Information Space*, available at: https://informationdemocracy.org/wp-content/uploads/2024/03/ID-AI-as-a-Public-Good-Feb-2024.pdf (accessed 28 August 2024).
15 Future of Life Institute, *Artificial Intelligence*, available at: https://futureoflife.org/cause-area/artificial-intelligence/ (accessed 28 August 2024).
16 Department for Science, Innovation & Technology (2023), *Emerging Processes for Frontier AI Safety,* available at: https://assets.publishing.service.gov.uk/media/653aabbd80884d000df71bdc/emerging-processes-frontier-ai-safety.pdf (accessed 28 August 2024).
17 Ministère de l'économie, des finances, et de la souveraineté industrielle et numérique, *Georges Akerlof*, available at: www.economie.gouv.fr/facileco/georges-akerlof (accessed 28 August 2024).

information asymmetry, a signal could be attached to a high-quality product, such as a certification, as explained by Akerlof. It is important to note that the EU AI Act foresees attaching signals, notably the CE marking, to high-risk AI systems to indicate regulatory conformity (Article 16 and Articles 43ff).

Even with full information symmetry, consumers might not choose responsible AI systems. The nudge theory developed by Sunstein and Thaler provides a helpful approach to understand what measures could be taken to promote the use of public interest AI systems. According to Sunstein and Thaler, a nudge alters people's behaviour without imposing restrictions. It encourages a certain behaviour and tries to make that behaviour more likely.[18] One could argue that the Fairtrade label works according to the nudge theory. It identifies ethical and fair products, thereby encouraging consumers to buy them without imposing these products. The FairPhone similarly tries to draw consumers' attention to the ethical and environmental impact of their phones, aiming to influence consumer behaviour. As further discussed in section 3.6 consumer awareness is still lacking regarding ethical AI, and needs to be fostered through education and awareness raising in parallel.

- ***A solution in the absence of regulation***

Even though some legislators are establishing rules for AI systems (EU AI Act, Brazilian draft AI bill, etc.) these do not concern AI systems worldwide. Thus, a voluntary certification mechanism could provide for a solution in contexts where regulations might not be the preferred option as they open doors to abusive authoritarian governments, but also in contexts where institutional and regulatory capacity and enforcement is weak or legislation is simply non-existent.[19] As a mechanism to be managed by an independent non-profit organisation (see Section 3 on governance), it would also be independent of governmental and geopolitical interests.

In addition, (in the European context, for example), this certification mechanism would also concern AI systems that are not as strictly regulated, notably those that are not characterised as high-risk and therefore do not need to undergo a third-party conformity assessment. Article 95 of the EU AI Act in fact encourages the elaboration of Codes of Conduct for voluntary application of specific requirements. In the same spirit as Article 95, a voluntary certification mechanism could present a solution to promote and identify responsible AI systems.

- ***Incentivizing the development of public interest AI***

As mentioned above, several countries have adopted / are adopting legislation on AI, and several international initiatives have developed AI principles and recommendations. Most initiatives highlight what a responsible AI system should look like, and propose measures to reduce risks of AI systems, the majority of which are built by privately owned companies. As Gerdes' analysis suggests, these systems work well for those that pay and not necessarily for those that are affected. Market competition also further encourages abandoning safety and ethics standards.[20] Therefore, the German government initiated a process to create a voluntary certification mechanism to assess the quality of AI systems, focusing on transparency, reliability, fairness, data protection and management, as well as cybersecurity.[21]

18 Thaler, R. H. and Sunstein, C. (2009), *Nudge: Improving Decisions About Health, Wealth, and Happiness*
19 Cafaggi, F., A. Renda and R. Schmidt (2013), «Transnational private regulation», in *International Regulatory Co-operation: Case Studies, Vol. 3: Transnational Private Regulation and Water Management*, OECD Publishing, Paris, https://doi.org/10.1787/9789264200524-3-en.
20 Kieslich, K. (2024), *The role of public opinion on ethical AI principles and its implication for a common good-oriented implementation*, University of Hohenheim, available at: https://hohpublica.uni-hohenheim.de/server/api/core/bitstreams/27902e30-8ee2-4ce8-83fa-08f92a16db87/content (accessed 28 August 2024).
21 Mission KI (2024), *Transparente Qualitäts- und Prüfstandards schaffen*, available at: https://mission-ki.de/ki-standards/ (accessed 4 September 2024).

Beyond an approach that is focused on quality, protecting fundamental rights, and reducing risks of AI systems, it is important to encourage the development of AI systems produced and deployed in the public interest. Such encouragement provides for an opportunity-based approach rather than solely a risk-reduction approach. A similar positive approach can be seen in the EU's Data Governance Act, which encourages data sharing for altruistic purposes.[22]

A certification mechanism would therefore aim to provide for a race to the top in the creation of public interest AI systems. By certifying public interest AI systems, the mechanism would also show that an alternative approach to AI development and deployment is possible. It would attach a competitive advantage to building public interest AI.

- ***Putting the definition of public interest in the hands of independent actors***

Self-regulation or industry-driven regulation, for example in the social media sphere, has proven to be insufficient. Even in cases where the democratic legislator has set the main rules regarding AI systems, vacuums on normative expectations remain and the defining of rules is transferred to standard-developing organisations or private actors.

For example, the EU AI Act foresees that the Commission makes standardisation requests to European standardisation organisations which are private entities (see Article 40). Even though the process foresees that all interested parties are included, it can be dominated by industry actors. The law also enables AI providers to opt for internal control as a procedure to assess conformity (Article 43), thereby giving the discretion to interpret conformity to the AI provider, even though this first-party conformity assessment is supposed to follow the same requirements.[23]

It is therefore important to set up a mechanism managed in the public interest, and not by private actors, which is responsible for defining the criteria of a public interest AI system.

- ***Responding to a transnational challenge and a fast-evolving issue***

As Cafaggi et al. explain, transnational private regulation responds to the specific challenge of transnational issues, where national regulation is insufficient and international regulation requires difficult-to-reach consensus.[24] The lack or shortcomings of global regulation explain, among other factors, the emergence of private regulation.[25] According to the authors, private regulation is also better-suited to respond to fast-evolving issues such as digital transformations.[26]

As AI represents a globally relevant and fast-evolving challenge, a certification mechanism could provide a valuable transnational alternative to the lack of international regulation. In addition, voluntary certification mechanisms can become de facto mandatory when integrated into other policy initiatives (for example public procurement, trade policy) and can also be integrated into legislation.

22 European Commission (2024), *Data Governance Act explained*, available at: https://digital-strategy.ec.europa.eu/en/policies/data-governance-act-explained (accessed 4 September 2024).
23 Casarosa, F. Cybersecurity certification of Artificial Intelligence: a missed opportunity to coordinate between the Artificial Intelligence Act and the Cybersecurity Act. *Int. Cybersecur. Law Rev.* 3, 115–130 (2022). https://doi.org/10.1365/s43439-021-00043-6
24 Cafaggi, F., A. Renda and R. Schmidt (2013), «Transnational private regulation», in *International Regulatory Co-operation: Case Studies, Vol. 3: Transnational Private Regulation and Water Management*, OECD Publishing, Paris, https://doi.org/10.1787/9789264200524-3-en.
25 Cafaggi, F. (2011), New Foundations of Transnational Private Regulation, *Journal of Law and Society*, vol. 38, no. 1, 2011, pp. 20–49, available at: www.law.berkeley.edu/wp-content/uploads/2019/10/Cafaggi-Foundations-of-TPR-2011.pdf (accessed 28 August 2024).
26 Cafaggi, F., A. Renda and R. Schmidt (2013), «Transnational private regulation», in *International Regulatory Co-operation: Case Studies, Vol. 3: Transnational Private Regulation and Water Management*, OECD Publishing, Paris, https://doi.org/10.1787/9789264200524-3-en.

1.3/ THE DENOMINATION OF THE MECHANISM – IDENTIFYING THE APPROPRIATE OBJECTIVE

The Forum's report suggests promoting "the creation of a tailored certification system for AI companies inspired by the success of the Fairtrade certification system".[27] While the report highlights the need for a certification mechanism, it does not provide for an in-depth discussion on the denomination of such a mechanism. The denomination, however, plays an important role as it provides for the scope and intention of the mechanism and will send a specific message.

Even though Fairtrade is a widely recognised and known concept, it primarily inspires and focuses on fair working conditions and pay, together with the trade of commodities.[28] It is also mainly used for commodity products, and it does not primarily focus on the concepts of the ethical, trustworthy, and responsible. The word "fair", however, does inspire fairness in the relation between two actors, and relates to ethics and trustworthiness. It is used by the initiative "Fairly Trained".[29]

A majority of initiatives focus on trustworthy, responsible, or ethical AI, such as those of the OECD, UNESCO, and the Global Index on Responsible AI.[30] The objective of the proposed certification mechanism is not only to set up a mechanism that identifies trustworthy, responsible, and ethical AI systems, but, in line with the Forum's report, one that inspires public interest: AI systems that are developed in the interest of society and align with values such as human rights, pluralism, diversity, and inclusion.

An in-depth discussion of the concept of public interest conducted by various scholars goes beyond the objective of this paper, yet it is helpful to define the main meaning of the concept. Drawing upon Eric R. Boot's analysis, the public interest is made up of two concepts, the public and its interest. The public in ancient Rome meant "pertaining to the concern of all the people", which is in contrast to the private, i.e., the interest of an individual or a group of individuals. Boot then takes a civic account of public interest, defining something as being in the public interest "if it increases the opportunities of the members of the public to pursue and realize the (permissible) ends they all share". The public interest is a widely used concept that is evoked in policymaking, and in law providing for public interest exceptions (for example in whistleblower protection).[31] The principle has also been applied to the media sector, which should serve the public interest and help create an informed citizenry. Protection of media freedom stems mainly from this perspective on the media which serves the public interest.[32]

Drawing upon this widely known concept, inspired by its impact on the media sector and the objective to promote the development and deployment of trustworthy, responsible, and ethical AI systems, this paper suggests calling the voluntary certification mechanism "public interest AI".

27 Forum on Information and Democracy (2024), *AI as a Public Good: Ensuring Democratic Control of AI in the Information Space*, available at: https://informationdemocracy.org/wp-content/uploads/2024/03/ID-AI-as-a-Public-Good-Feb-2024.pdf (accessed 28 August 2024).
28 Max Havelaar Stiftung, , *Qu'est ce que Fairtrade?*, available at: www.fairtrademaxhavelaar.ch/fr/quest-ce-que-fairtrade (accessed 28 August 2024).
29 Fairly Trained, available at: www.fairlytrained.org/ (accessed 28 August 2024).
30 Global Index on Responsible AI (2024), available at: www.global-index.ai/methodology (accessed 3 September 2024).
31 Boot, E.R. (2024), The Public Interest: Clarifying a Legal Concept. *Ratio Juris*, 37: 110-129. https://doi.org/10.1111/raju.12401
32 Croteau, D. and Hoynes, W. (2006), *The business of media: corporate media and the public interest*, Pinge Forge Press

1.4/ DEFINING THE SCOPE OF THE CERTIFICATION MECHANISM

AI systems are being deployed at all levels of society, government, and the economy, including business-to-business and business-to-consumer products. The development of AI products is based on a complex value chain. Thus, it is important to identify the scope of the voluntary certification mechanism.

While some international regulations or guiding documents, as the Council of Europe's Framework Convention apply to the whole lifecycle of AI systems used by "public authorities, or private actors acting on their behalf",[33] regardless of the type of AI systems, others, as the EU's AI Act differentiate the regulatory obligations depending on the type of AI system (in this case categorised by its risk). Similarly, most existing voluntary certification mechanisms for products are not an all-encompassing assessment but focus on very specific issues and parts of the value chain.

In this context, the Fairtrade certification guarantees that the original product has been produced under fair conditions (fair pay, etc). It does not certify that the whole value chain of the product (i.e., from the production to the consumer passing through retailers) adheres to the values of Fairtrade. Fairly Trained also only assesses if the AI developer fairly obtained its training data, without assessing the whole data production chain.

According to the experts interviewed during this research, it is important to clearly define the objectives to be achieved and the problems to be solved. Instead of an all-encompassing certification mechanism with a wide scope, the OECD recommends niche standards. While these can be linked to international standards, their added value comes from addressing a single specific issue.

The objective of this paper is to explore the creation of a certification mechanism that can assess the public interest character of an AI system. Therefore, such a mechanism does not concern all AI systems and their entire lifecycle, but rather those that are directly accessible to users. These include mainly large language models and artificial general intelligence.

The AI developers and providers are involved in building AI systems that are accessible to the users. While the AI developer builds the foundational model, the provider[34] adapts it and fine-tunes it for deployment. As the certification system aims to identify public interest AI systems available to users, the paper suggests focusing the certification mechanism on the provider's AI system. However, in addition, the elaboration of a voluntary certification mechanism for foundation models could be envisioned, which would then guide AI providers in their choice.

33 Article 3, Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law https://rm.coe.int/1680afae3c

34 *AI "provider' means a natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge*; as defined in Article 3 of the EU AI Act

2. DEFINING THE CERTIFICATION CRITERIA

2.1/ INTRODUCTION AND APPROACH

The certification mechanism suggested in this paper does not concern all AI systems available, but only those directly available to users. Furthermore, it focuses only on AI providers and not AI developers. Therefore, the mechanism's criteria of assessment should also align clearly with the objective of identifying what constitutes a public interest system. As such, the certification mechanism would not cover all possible criteria, but only those that primarily define the public interest character of the AI system. These criteria will be discussed in this section.

The establishment of the proposed certification mechanism does not exclude the creation of further labels that assess other aspects of the AI system or the development of secondary criteria, which are not obligatory. The Fairtrade certification, for example, includes core criteria which are mandatory, and development requirements, which are more aspirational.[35] Furthermore, while the following discussion includes an initial suggestion for the criteria to be included, the certification mechanism and its criteria will have to evolve in line with legislative and normative developments and latest development of AI systems and knowledge.

In addition to defining the criteria to be assessed, it is necessary to define the grading system. Certification and assessment mechanisms take different approaches to grading the product. Many certification mechanisms provide a binary grade: the product is or is not fairtrade, the product is or is not fairly trained, the media satisfies or does not satisfy the Journalism Trust Initiative (JTI) criteria. Others, such as the Nutriscore, or several rankings of digital systems (Ranking Digital Rights[36] and the Foundation Model Transparency Index[37]) rank products on a scale. The Forest Stewardship Council (FSC), which also builds upon a binary grade, introduced different labels to take into account the different typology of products: "FSC 100%", "FSC Recycled", "FSC Mixed".[38]

The suggested certification mechanism does not aim only to assess the public interest character of an AI system, but also to encourage the development and use of such systems through other policies (public procurement, etc. see Section 3.6). Hence, it is advisable to opt for a binary grading system. Such a system would also provide for an easily identifiable user signal.

With this approach in mind, the following section discusses the criteria that should be assessed by a certification mechanism for public interest AI. Inspiration for these criteria is taken from existing international principles, such as those of the OECD,[39] UNESCO,[40] the Council of Europe,[41] and the

35 Fairtrade International, *Fairtrade Standards*, available at: www.fairtrade.net/standard (accessed 29 August 2024).
36 Ranking Digital Rights, available at: https://rankingdigitalrights.org/ (accessed 29 August 2024).
37 Miller, K. (2023), Introducing *The Foundation Model Transparency Index,* Stanford University Human-Centered Artificial Intelligence, available at: https://hai.stanford.edu/news/introducing-foundation-model-transparency-index, (accessed 28 August 2024).
38 FSC, *What's in a label?* accessible at: https://fsc.org/en/what-the-fsc-labels-mean (accessed 29 August 2024).
39 Including the OECD (2023), *OECD Guidelines for Multinational Enterprises on Responsible Business Conduct*, OECD Publishing, Paris, https://doi.org/10.1787/81f92357-en. and the forthcoming, *OECD Due Diligence for Responsible AI*
40 UNESCO (2022), *Recommendation on the Ethics of Artificial Intelligence*, available at https://unesdoc.unesco.org/ark:/48223/pf0000381137 (accessed 28 August 2024).
41 Council of Europe (2024), *Council of Europe Framework Convention on Artificial Intelligence and human rights, democracy and the rule of law*, available at: https://rm.coe.int/1680afae3c

Hiroshima Code of Conduct,[42] together with policy analysis and recommendations from the Forum's report, *The Global Index on Responsible AI*,[43] and the Paris Charter on AI and Journalism,[44] as well as risk management frameworks such as the NIST's *Trustworthy and Responsible AI*,[45] and ISO/IEC 23894:2023.[46]

2.2/ CRITERIA FOR ASSESSING THE COMPANY OR ENTITY PROVIDING THE AI SYSTEM

As Boot[47] discusses, the public interest is often evoked without a clear definition of what it is nor how it can be determined, leaving such interpretation to courts or policymakers. Therefore, this paper suggests several criteria that can be used as proxies to identify if an AI system is in the public interest. The first set of criteria concerns the entity deploying the AI system. The provider's adherence to and respect for democratic values and fundamental rights is a first indication of the public interest character of the AI system. This is in line with ISO/IEC 17065 which includes requirements not only for the object (e.g., the AI system) but also for the provider (e.g., the AI provider) to establish certification.

- ***Official commitment to human rights***

All key policy and legislative documents on AI put the respect for, protection, and promotion of human rights in the AI lifecycle at their core, such as the UNESCO's Ethics Principles,[48] the OECD AI Principles,[49] and the Council of Europe Convention.[50] An in-depth assessment of compliance is a complex matter. However, as a first step, as suggested by the OECD *Guidelines for Multinational Enterprises*, companies should at minimum have a "publicly available policy commitment to respect human rights" and "carry out human rights due diligence".[51] Even though an official commitment is no guarantee for implementation of that commitment, it is an important baseline. An official commitment to human rights and human rights due diligence could also be applied to AI providers.

- ***Democratic, inclusive and transparent governance of the AI provider***

As the Forum's report[52] discusses, democratic governance and transparency of the AI provider aim to ensure that AI development and deployment is not guided by private interests and shareholders, but pursued in the public interest. In this respect, transparency is an essential element of control by the users, and democratic governance is an element to provide a space for public interest actors in the decision-making processes. In addition, inclusiveness and diversity of teams are important to ensure

42 Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems, available at: www.mofa.go.jp/files/100573473.pdf
43 Global Index on Responsible AI (2024), available at: www.global-index.ai/methodology (accessed 3 September 2024).
44 RSF (2023), *Paris Charter on AI and Journalism*, available at: https://rsf.org/sites/default/files/medias/file/2023/11/Paris%20Charter%20on%20AI%20and%20Journalism.pdf (accessed 28 August 2024).
45 NIST Trustworthy and Responsible AI, *NIST AI 600-1, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*,, available at : https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf (accessed 28 August 2024).
46 ISO/IEC 23894:2023, Information technology — Artificial intelligence — Guidance on risk management, available at: www.iso.org/standard/77304.html (accessed 28 August 2024).
47 Boot, E.R. (2024), The Public Interest: Clarifying a Legal Concept. *Ratio Juris*, 37: 110-129. https://doi.org/10.1111/raju.12401
48 UNESCO (2022), *Recommendation on the Ethics of Artificial Intelligence*, available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137 (accessed 28 August 2024).
49 OECD (2024), *Recommendation of the Council on Artificial Intelligence*, available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (accessed 29 August 2024).
50 Council of Europe (2024), *Council of Europe Framework Convention on Artificial Intelligence and human rights, democracy and the rule of law*, available at: https://rm.coe.int/1680afae3c
51 OECD (2023), *OECD Guidelines for Multinational Enterprises on Responsible Business Conduct*, OECD Publishing, Paris, https://doi.org/10.1787/81f92357-en.
52 Forum on Information and Democracy (2024), *AI as a Public Good: Ensuring Democratic Control of AI in the Information Space*, available at: https://informationdemocracy.org/wp-content/uploads/2024/03/ID-AI-as-a-Public-Good-Feb-2024.pdf (accessed 28 August 2024).

that AI systems work for all and are non-discriminatory. Finally, whistleblower protection is an important safeguard against internal wrongdoing and risks.

As recognised by the OECD *Guidelines for Multinational Enterprises*, but also voluntary certification mechanisms like JTI,[53] transparency is of particular concern in the business model and ownership. The business model and financing schemes provide valuable insights regarding the interests that guide the AI provider. Disclosure of ownership and beneficial ownership enable user control and transparency regarding private interests. Transparency is also a sign of commitment to the users' right to know.

Transparency is a minimum requirement to assess the AI provider's business interests. Yet, the public's interest can only be guaranteed through democratic governance and oversight structures. Therefore, in line with UNESCO's principles, different stakeholders should be involved in the governance structure of the AI provider, notably public interest representatives.[54] This involvement can take various forms, such as civil society or researcher representatives and/or user representatives with seats either on the board, or on an advisory board with at least some kind of decision-making power. It is important that all key user groups are represented in this governance framework, which means – in the case of multinational enterprises – that user representatives from different regions and their needs should be taken into account.

As public interest AI systems will serve the public, safeguards need to be in place to fight against discrimination and bias in AI systems that can originate from training data, annotation, and modelling choices. Studies have shown that AI developers are not aware of all ethical choices they face and all the potential discriminatory outcomes AI systems can produce.[55] Consequently, a first indicator to assess if an AI provider takes diversity and non-discrimination seriously is the set-up of the teams that build the AI. As highlighted in the NIST framework, AI entities should "establish and empower interdisciplinary teams that reflect a wide range of capabilities, competencies, demographic groups, domain expertise, educational backgrounds, lived experiences, professions, and skills across the enterprise to inform and conduct risk measurement and management functions".[56] Diversity in terms of potentially discriminated categories such as "race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status"[57] must also be reflected in the teams.

A final essential element of democratic and transparent governance of the AI provider is whistleblower protection or a right to warn, as called for by current and former OpenAI employees.[58] The NIST framework,[59] OECD's *Guidelines for Multinational Enterprises*,[60] as well as the Forum's report[61] recognise the importance of whistleblower protection for current and former employees. Whistleblower protection encourages these employees to provide valuable insights on risks and wrongdoings that can endanger the public interest character of an AI system.

53 JTI, CWA No. 17493, *CEN Workshop Agreement published on December 19th, 2019*, available at: www.jti-app.com/footer/cwa#standard_clauses (accessed 28 August 2024).
54 Forum on Information and Democracy (2024), *AI as a Public Good: Ensuring Democratic Control of AI in the Information Space*, available at: https://informationdemocracy.org/wp-content/uploads/2024/03/ID-AI-as-a-Public-Good-Feb-2024.pdf (accessed 28 August 2024).
55 Kieslich, K. (2024), *The role of public opinion on ethical AI principles and its implication for a common good-oriented implementation*, University of Hohenheim, available at: https://hohpublica.uni-hohenheim.de/server/api/core/bitstreams/27902e30-8ee2-4ce8-83fa-08f92a16db87/content (accessed 28 August 2024).
56 NIST Trustworthy and Responsible AI, *NIST AI 600-1, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*,, available at : https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf (accessed 28 August 2024).
57 Article 2 - Universal Declaration of Human Rights, www.un.org/en/about-us/universal-declaration-of-human-rights
58 Samuel, S. (2024),OpenAI insiders are demanding a "right to warn" the public, *Vox*, available at: www.vox.com/future-perfect/353933/openai-open-letter-safety-whistleblowers-right-to-warn (accessed 28 August 2024).
59 NIST Trustworthy and Responsible AI, *NIST AI 600-1, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*, available at : https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf (accessed 28 August 2024).
60 OECD (2023), *OECD Guidelines for Multinational Enterprises on Responsible Business Conduct*, OECD Publishing, Paris, https://doi.org/10.1787/81f92357-en.
61 Forum on Information and Democracy (2024), *AI as a Public Good: Ensuring Democratic Control of AI in the Information Space*, available at: https://informationdemocracy.org/wp-content/uploads/2024/03/ID-AI-as-a-Public-Good-Feb-2024.pdf (accessed 28 August 2024).

- ***Employment and working conditions***

Various "workers" are involved in building and developing AI systems. These range from the IT and computer scientists employed within the AI provider to outsourced data-labellers. The working conditions of all these actors are another indicator of respect for the public interest by the AI provider, as well as their respect for human rights, particularly Article 23 of the Universal Declaration of Human Rights (UDHR). Employment conditions are at the core of several sustainability certification mechanisms taken as inspiration for this paper, particularly Fairtrade, but also FSC. Employment conditions are also taken into account by the Global Index on Responsible AI and JTI. The OECD AI principles explicitly call for the respect of internationally recognised labour rights by AI actors.

An AI provider should at minimum uphold the obligations and commitments of the International Labour Organization (ILO) Declaration on Fundamental Principles and Rights at Work, namely:

- "freedom of association and the effective recognition of the right to collective bargaining;
- the elimination of all forms of forced or compulsory labour;
- the effective abolition of child labour;
- the elimination of discrimination in respect of employment and occupation; and
- a safe and healthy working environment".[62]

In addition, the UDHR provides for the right to just and favourable remuneration (Article 23.3). Fair remuneration is of particular importance in the AI context, where the exploitation of outsourced data workers has been criticised.[63]

- ***Environmental sustainability***

An AI system cannot be considered to be a public interest AI system if it is not aligned with the values of environmental sustainability, as the preservation of the planet is a global goal.

The UNESCO principles enshrine the value of "environmental and ecosystem flourishing" for AI systems.[64] The EU AI Act calls for AI systems to be developed and used in "a sustainable and environmentally friendly manner as well as in a way to benefit all human beings".[65] As identified by UNESCO, the "carbon footprint, energy consumption and the environmental impact of raw material extraction"[66] are of particular concern for AI systems.

The Blue Angel ecolabel for software products can provide an inspiration for the development of criteria to assess the environmental sustainability of digital products. The label includes transparency about the energy consumption of the software, low-enough performance requirements to run the software on older hardware, and user autonomy in defining its environmental impact.[67] This focus on transparency is reiterated in the NIST framework and the UNESCO's recommendations, which proposes documenting anticipated environmental impacts of the AI system throughout its lifecycle. Hence, to assess an AI

62 ILO (2022), *ILO Declaration on Fundamental Principles and Rights at Work*, available at: www.ilo.org/ilo-declaration-fundamental-principles-and-rights-work (accessed 28 August 2024).
63 Rowe, N. (2023), Millions of Workers Are Training AI Models for Pennies, *Wired*, available at: www.wired.com/story/millions-of-workers-are-training-ai-models-for-pennies (accessed 28 August 2024).
64 UNESCO (2022), *Recommendation on the Ethics of Artificial Intelligence*, available at https://unesdoc.unesco.org/ark:/48223/pf0000381137 (accessed 28 August 2024).
65 EU (2024), *Artificial Intelligence Act*, available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689&qid=1722524514731
66 UNESCO (2022), *Recommendation on the Ethics of Artificial Intelligence*, available at https://unesdoc.unesco.org/ark:/48223/pf0000381137 (accessed 28 August 2024).
67 Blue Angel (2020), *Resource and Energy-Efficient Software Products*, Basic Award Criteria Edition January 2020 Version 3, available at: https://produktinfo.blauer-engel.de/uploads/criteriafile/en/DE-UZ%20215-202001-en-Criteria-V3.pdf (accessed 28 August 2024).

system's respect for environmental sustainability, the availability and disclosure of an environmental impact assessment could provide useful criteria for assessment.

In addition to transparency about the environmental impact, an AI provider must put in place measures to reduce the negative effects of its AI system to align its practices with public interest values. As the impact of AI systems on the environment (negative effects on ecosystems through resource extraction, carbon emissions, energy consumption, e-waste, etc.) as well as the appropriate solutions vary according to the AI system, it is advisable rather to opt for assessment criteria regarding the processes in place, than the actual measures to reduce the environmental impact.

The EU AI Act, for example, foresees that AI entities will need to report serious incidents to the environment to relevant authorities. The Act further established that a voluntary Code of Conduct could enumerate further requirements for environmental sustainability (Article 95).

The processes in place to reduce the environmental impact should include the assessment of environmental impacts, the development and implementation of solutions, measuring progress and impact of solutions, and reporting incidents. Transparency and democratic governance should govern these processes. As UNESCO highlights, an inclusive process should in particular involve local and indigenous communities, as they might be most impacted by environmental consequences.[68]

SUMMARY OF MINIMUM CRITERIA TO ASSESS THE AI PROVIDER OF THE PUBLIC INTEREST AI SYSTEM

Official commitment to human rights	• Publicly available policy commitment to respect human rights • Human rights due diligence
Democratic and transparent governance	• Transparency of business model and financing • Transparency of ownership and beneficial ownership • Democratic governance with public interest representatives with decision-making power • Diverse and inclusive teams • Whistleblower protection
Employment and working conditions	• Respect ILO Declaration on Fundamental Principles and Rights at Work • Just and fair remuneration for all "workers" involved in the AI system development
Environmental sustainability	• Publicly available assessment of the environmental impact of the AI system throughout its whole life cycle • Transparent and democratic processes including local actors to assess the environmental impact, develop solutions, measure progress and report incidents

The OECD *Guidelines for Multinational Enterprises on Responsible Business Conduct* show that further elements could be considered as relevant to identify responsible business conduct of a company, such as respect for anti-bribery and corruption legislation, obeying domestic laws, respecting tax obligations, respecting consumer rights, and providing relevant information.[69] Respect for these guidelines or the elements contained in these guidelines could be secondary criteria or figure in additional labels. Further reflections are needed to assess if any of these criteria are essential to assess if an AI provider works in the public interest, bearing in mind that the certification mechanism should be focused enough to achieve its objective.

68 UNESCO (2022), *Recommendation on the Ethics of Artificial Intelligence*, available at https://unesdoc.unesco.org/ark:/48223/pf0000381137 (accessed 28 August 2024).
69 OECD (2023), *OECD Guidelines for Multinational Enterprises on Responsible Business Conduct*, OECD Publishing, Paris, https://doi.org/10.1787/81f92357-en.

2.3/ THE AI SYSTEM AND ITS CHARACTERISTICS

The internal governance and values of the AI provider offers a proxy for the public interest character of the AI system, however there are also elements of the system itself that need to be assessed to identify if it was developed in, and is supposed to serve, the public interest.

- ***Public interest objective***

A first and logical criterion is, therefore, the objective of the AI system, its expected and foreseen use. The OECD AI principles call for trustworthy AI in "pursuit of beneficial outcomes for people and the planet"[70] and those of UNESCO for ensuring "peaceful and just societies, which is based on an interconnected future for the benefit of all, consistent with human rights and fundamental freedoms".[71] The foreseen and expected uses of AI should therefore serve the public interest. To better evaluate the truthfulness of the foreseen objective, AI entities should also make available the value choices that served as the basis to build and adapt the AI system.

- ***Fairness, privacy and non-discrimination***

Fairness refers to being reasonable, right, and just, but also equal treatment.[72] The OECD AI principles highlight fairness as one of the key values AI systems should respect.[73] As the initiative Fairly Trained highlights, an AI system can only be considered fair if it respects the legitimate rights of data creators. Similarly, a product cannot be considered fairly traded if it does not respect the rights of its producers, such as fair wage and fair working conditions. Fairness is closely linked with the public interest, as one cannot pretend to build a public interest AI system without treating data creators fairly.

The question thus arises of what is considered fair and in the public interest in terms of data curation and use. This concerns the respect for copyright and compensation. Fairly Trained has a very strict approach stating that data sources are only considered to be used fairly if there is a contractual agreement, if they are available under an open licence, or in the public domain globally. Others consider that invoking "fair use" can be sufficient, instead of consent or a contractual agreement.[74] In any case, at very least the data sources need to be disclosed in a searchable database[75] to enable any data creators to verify the respect for rights and the users to understand upon what kind of data sources the system was built.

Respect for data protection and privacy are not only reiterated in all major guidelines and legislation regarding AI (Council of Europe, OECD, UNESCO), but are a human right. Therefore, such respect is also essential for a public interest AI system, which should reflect the common ends society shares, within permissible bounds. It follows that AI systems must not only respect existing data protection laws in data curation and use, but also implement data minimization and privacy by design, and provide for meaningful choices to opt-out of tracking and personalisation. These elements are also included in UNESCO's *Ethical Impact Assessment*.[76] As the NIST framework outlines, this includes providing "human subjects with options to withdraw participation or revoke their consent for present or future use of their

70 OECD (2024), *Recommendation of the Council on Artificial Intelligence*, available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (accessed 29 August 2024)
71 UNESCO (2022), *Recommendation on the Ethics of Artificial Intelligence*, available at https://unesdoc.unesco.org/ark:/48223/pf0000381137 (accessed 28 August 2024).
72 Collins, *Definition of 'fairness'*, available at: https://www.collinsdictionary.com/dictionary/english/fairness (accessed 28 August 2024).
73 OECD (2024), *Recommendation of the Council on Artificial Intelligence*, available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (accessed 29 August 2024).
74 The issue of fair use of data sources is a central element to public interest AI. Yet, given the ongoing discussions in several international fora and the complexity of the issue, the paper does not provide for a conclusive answer.
75 Refer to page 27 of the Forum report for a detailed discussion what this database should include
76 UNESCO (2023), *Ethical impact assessment: a tool of the Recommendation on the Ethics of Artificial Intelligence*, available at: https://unesdoc.unesco.org/ark:/48223/pf0000386276 (accessed 3 September 2024).

data in GAI applications".[77] However, as several legal texts point out (EU GDPR & UK GDPR) under certain conditions the public interest can be invoked to process special category data.[78] In this regard, the EU AI Act creates specific conditions for the processing of personal data to build AI systems in the public interest (Article 59).

Finally, an AI system can only serve all the people if it is non-discriminatory and respects equality and diversity, values recognized internationally as core elements of AI systems (Council of Europe, OECD, UNESCO). The Global Index for Responsible AI therefore highlights the importance of tackling bias and unfair discrimination and respecting cultural and linguistic diversity.[79] The UNESCO *Ethical Impact Assessment* suggests assessing the AI's potential discriminatory outcomes and testing data against bias.[80] As the Forum report suggests, AI providers should conduct impact assessments of their system prior to deployment, and review them regularly to check for unfair and reinforcing bias, discrimination, and inaccuracies and misrepresentation in different languages and cultures. These assessments should be public to allow for user control.[81]

- ***Transparency of the system and external scrutiny***

Transparency of the AI system enables external scrutiny and verification of how the system was built and works. As in the field of democratic governance, where government transparency serves to enable citizens to verify if the government / public institution works in the public interest, transparency of AI systems serves the same purpose. Transparency features in all key AI guidelines and legislation. The Forum's report includes an in-depth discussion of all AI system information that should be made available to the general public and the more detailed information that should be available to regulators (see Chapter 4.4).[82] The Stanford University Foundation Model Transparency Index[83] already provides for a useful assessment of AI systems' transparency. The level of transparency of an AI system would on the one hand be very insightful to assess the AI provider's willingness to be open to inspection, yet it is also a very complex assessment. Further discussion is needed to evaluate the usefulness and feasibility of such a criterion.

However, criteria focusing on access for independent vetted researchers and regulators to study the AI system in detail could be envisioned. This kind of access may eventually be possible, in certain cases, under the Digital Services Act, but currently the closest example seems to be the agreement between the US Safety Institute and OpenAI and Anthropic, enabling this government institute to conduct external evaluation of the companies' AI systems.[84] External access would enable a full and in-depth study and scrutiny of AI systems, enabling external stakeholders to verify the system's public interest character. The Forum's report already calls for such access; the UK's Department for Science, Innovation and Technology guidelines for frontier AI systems equally recognises the importance of allowing independent and external evaluators to conduct model evaluations.[85] Similarly, UNESCO's *Ethical Impact*

77 NIST Trustworthy and Responsible AI, *NIST AI 600-1, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*, available at : https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf (accessed 28 August 2024).
78 ICO, *What are the substantial public interest conditions?*, available at: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/special-category-data/what-are-the-substantial-public-interest-conditions/ (accessed 28 August 2024).
79 Global Index on Responsible AI (2024), available at: www.global-index.ai/methodology (accessed 3 September 2024).
80 UNESCO (2023), *Ethical impact assessment: a tool of the Recommendation on the Ethics of Artificial Intelligence*, available at: https://unesdoc.unesco.org/ark:/48223/pf0000386276 (accessed 3 September 2024).
81 Forum on Information and Democracy (2024), *AI as a Public Good: Ensuring Democratic Control of AI in the Information Space*, available at: https://informationdemocracy.org/wp-content/uploads/2024/03/ID-AI-as-a-Public-Good-Feb-2024.pdf (accessed 28 August 2024).
82 Forum on Information and Democracy (2024), *AI as a Public Good: Ensuring Democratic Control of AI in the Information Space*, available at: https://informationdemocracy.org/wp-content/uploads/2024/03/ID-AI-as-a-Public-Good-Feb-2024.pdf (accessed 28 August 2024).
83 Miller, K. (2023), *Introducing The Foundation Model Transparency Index*, Stanford University Human-Centered Artificial Intelligence, available at: https://hai.stanford.edu/news/introducing-foundation-model-transparency-index, (accessed 28 August 2024).
84 NIST (2024), *U.S. AI Safety Institute Signs Agreements Regarding AI Safety Research, Testing and Evaluation With Anthropic and OpenAI*, available at: www.nist.gov/news-events/news/2024/08/us-ai-safety-institute-signs-agreements-regarding-ai-safety-research (accessed 3 September 2024).
85 Department for Science, Innovation & Technology (2023), *Emerging Processes for Frontier AI Safety*, available at: https://assets.publishing.service.gov.uk/media/653aabbd80884d000df71bdc/emerging-processes-frontier-ai-safety.pdf (accessed 28 August 2024).

Assessment focuses on algorithmic explainability through openness to the public or oversight institutions or open source.[86]

- ***Inclusive and participatory AI system development***

The development of an AI system, as well as its risk management, red-teaming, and fine-tuning processes involve value-laden choices, for example regarding data curation or acceptable and unacceptable risks or biases. Thus, to ensure that an AI system is built in the public interest, these choices cannot be taken by the AI provider behind closed doors. In line with the society-in-the loop framework developed by Rahwan,[87] AI providers need to put in place processes for meaningful and sustained inclusive and participatory AI system development. Similarly, the draft OECD *Due Diligence Guidance for Responsible AI* puts an emphasis on stakeholder engagement. The certification mechanism could therefore include criteria on the involvement of different stakeholders, such as independent civil society, researchers, media, and affected communities, ensuring representativeness and diversity. Equally, this criterion should include elements on the participation process itself (i.e., at which stages of the AI development process stakeholders are involved and with which decision-making power).

An inclusive and participatory process aims to tailor AI systems as much as possible to the public interest and reduce any potential risks. However, issues might arise or become evident post-deployment. The Forum's report therefore suggests the implementation of user and public feedback mechanisms.[88] Similarly, the NIST framework calls for the establishment of user feedback mechanisms.[89] Criteria should therefore assess the existence of this kind of mechanism as well as its usefulness and impact.

SUMMARY OF MINIMUM CRITERIA TO ASSESS THE AI SYSTEM'S PUBLIC INTEREST CHARACTER

Public interest objective of the system	• Expected and foreseen use of the AI system is in the public interest • Transparency about the value choices to build the AI system
Fairness, privacy & non-discrimination	• Fair treatment of data sources and creators • Disclosure of data sources • Respect for data protection laws & data minimization • Meaningful choice to opt out of tracking and personalisation • Publicly available and regular impact assessments to assess bias, discrimination and inaccuracies in different cultures and languages
Transparency & external scrutiny	• Regulator and independent vetted researcher access to the AI system to study it from within
Inclusive & participatory AI system development	• Existence of meaningful inclusive and participatory processes to develop the AI system • Existence of feedback mechanisms post-deployment to involve stakeholders

86 UNESCO (2023), *Ethical impact assessment: a tool of the Recommendation on the Ethics of Artificial Intelligence*, available at: https://unesdoc.unesco.org/ark:/48223/pf0000386276 (accessed 3 September 2024).
87 Kieslich, K. (2024), *The role of public opinion on ethical AI principles and its implication for a common good-oriented implementation*, University of Hohenheim, available at: https://hohpublica.uni-hohenheim.de/server/api/core/bitstreams/27902e30-8ee2-4ce8-83fa-08f92a16db87/content (accessed 28 August 2024).
88 Forum on Information and Democracy (2024), *AI as a Public Good: Ensuring Democratic Control of AI in the Information Space*, available at: https://informationdemocracy.org/wp-content/uploads/2024/03/ID-AI-as-a-Public-Good-Feb-2024.pdf (accessed 28 August 2024).
89 NIST Trustworthy and Responsible AI, N*IST AI 600-1, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*, available at : https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf (accessed 28 August 2024).

Several other issues are regularly included in the requirements for responsible AI, such as the safety and cybersecurity of the system as well as the risk assessment and mitigation measures put in place, for example key issues of the Hiroshima Code of Conduct and of the UNESCO *Ethical Impact Assessment*. The *Ethical Impact Assessment* also highlights the importance of human oversight.[90] These elements are crucial in assessing the risks of an AI system. However, in the interest of developing a "niche" certification mechanism, and considering that risk management and security practices evolve constantly, this paper suggests excluding criteria on these topics from the certification mechanism. It further argues that inclusive and participatory processes for the development of AI systems already ensure that decisions on safety and risk management are taken from a public interest point of view.

90 Department for Science, Innovation & Technology (2023), *Emerging Processes for Frontier AI Safety*, available at: https://assets.publishing.service.gov.uk/media/653aabbd80884d000df71bdc/emerging-processes-frontier-ai-safety.pdf (accessed 28 August 2024). UNESCO (2023), *Ethical impact assessment: a tool of the Recommendation on the Ethics of Artificial Intelligence,* available at: https://unesdoc.unesco.org/ark:/48223/pf0000386276 (accessed 3 September 2024).

3. GOVERNANCE OF THE CERTIFICATION MECHANISM

The governance of a certification mechanism and the status and type of scheme owner[91] are as important as the criteria it assesses, as it impacts the credibility of the mechanism and therefore its potential uptake and impact. The International Social and Environmental Accreditation and Labelling Alliance (ISEAL Alliance), a UK-registered Charity with members worldwide, has developed Credibility Principles[92] to define what makes a sustainability system credible.[93]

Their Code of Good Practice[94] outlines all the requirements a credible system should fulfil. These principles and requirements can be taken as a benchmark to develop the governance scheme of a public interest AI certification mechanism and will be referenced throughout this paper to provide suggestions.

Overview of the ISEAL Credibility Principles:

- ***Sustainability impacts:*** A credible sustainability system makes a difference where it matters.
- ***Collaboration:*** A credible sustainability system works with others to create change.
- ***Value creation:*** A credible sustainability system adds value.
- ***Measurable progress:*** A credible sustainability system can demonstrate the difference it is making.
- ***Stakeholder engagement:*** A credible sustainability system listens and learns.
- ***Transparency:*** A credible sustainability system earns trust by being open and honest.
- ***Impartiality:*** A credible sustainability system is impartial.
- ***Reliability:*** A credible sustainability system provides trustworthy assessments of users' performance.
- ***Truthfulness:*** A credible sustainability system's claims and communications can be trusted.
- ***Continual improvement:*** A credible sustainability system keeps improving.[95]

In addition, the OECD's work on due diligence provides insights for what makes a credible mechanism. It highlights the importance of an assessment methodology which not only evaluates the written standards of the product, but also its actual implementation. The OECD has also developed a methodology which assesses a certification mechanism against OECD due diligence principles, enabling the mechanism to align itself with existing internationally recognised standards.[96] This is of particular relevance for a public

91 According to NIST, the scheme owner is the "person or organization responsible for developing and maintaining a specific certification scheme". Available at: www.nist.gov/glossary-term/31586 (accessed 3 September 2024).
92 ISEAL, *ISEAL Credibility Principles*, available at: www.isealalliance.org/defining-credible-practice/iseal-credibility-principles (accessed 28 August 2024).
93 Vermeulen, W.J.V et al (2010), *Roles of Governments in Multi-Actor Sustainable Supply Chain Governance Systems and the effectiveness of their interventions: An Exploratory Study*, Netherlands Environmental Assessment Agency (PBL), The Hague/Bilthoven, available at: www.pbl.nl/sites/default/files/downloads/PBL-2010-Other-roles-for-governments-needed-500411001.pdf (accessed 29 August 2024).
94 ISEAL (2023), *ISEAL Code of Good Practice for Sustainability Systems*, available at: www.isealalliance.org/sites/default/files/resource/2024-03/ISEAL-Code-of-Good-Practice-v1.0_ISEAL_final_public.pdf.pdf (accessed 28 August 2024).
95 ISEAL, *ISEAL Credibility Principles,* available at: www.isealalliance.org/defining-credible-practice/iseal-credibility-principles (accessed 28 August 2024).
96 OECD (2022), *The role of sustainability initiatives in mandatory due diligence: Background note on Regulatory Developments concerning Due Diligence for Responsible Business Conduct*, available at : https://mneguidelines.oecd.org/the-role-of-sustainability-initiatives-in-mandatory-due-diligence-note-for-policy-makers.pdf (accessed 28 August 2024).

interest AI certification mechanism, which could be aligned with the OECD *Due Diligence Guidance for Responsible Business Conduct*[97] and the forthcoming OECD *Due Diligence Guidance for Responsible AI.*

Finally, the governance of the certification mechanism should respect internationally recognised standards on certification, such as the ISO/IEC 17065 - Conformity assessment. This standard sets requirements for bodies certifying products, processes and services including requirements for the certification processes they conduct (i.e., evaluation, review, certification decision, and surveillance – the assessment methodology).[98] This standard has been recognised worldwide and has been adopted by many certification mechanisms including FSC and JTI. Building upon this standard would provide for a robust and more credible process to certify public interest AI systems.

3.1/ STANDARD SETTING

One of the key elements of a certification mechanism is to set a standard,[99] which in line with the ISEAL Credibility Principles[100] should make a difference (the sustainability impacts principle). In line with these principles, such a standard should be established through a collaborative process and stakeholder engagement, providing all concerned parties with the possibility of influencing the standard.

The certification mechanisms examined for this paper have taken different approaches to setting standards. This includes differences in the process of how to set and revise the standards, but also differences regarding what the standards are built on. All have adopted inclusive and collaborative approaches, some more and others less structured.

Fairly Trained conducted informal consultations with relevant stakeholders, while more established mechanisms like Fairtrade and FSC have clear policies on standard setting and revision.[101] These policies include open consultations and decisions by the governance board. In contrast to these initiatives, in which organisations followed their own policies, the Journalism Trust Initiative worked with the European Committee for Standardization (CEN) to develop the criteria for its initiative. In doing this, they followed a standardised procedure which provided for legitimacy and clear rules on revision. The International Organization for Standardization (ISO) also provides a mechanism to set international standards.

Several of these standard developing organisations, such as ISO, IEC, IEEE, IETF, CEN and CENELEC have already started developing standards on artificial intelligence. A joint committee on AI of ISO and IEC (ISO/IEC JTC 1/SC 42) looks into conceptual standards, risk management, identifying bias, and collecting and referencing AI use cases, while European Standardisation Organisations (ESOs) are working on standards required by the European AI Act. Cantero Gamito & Marsden criticise the excessive reliance upon these organisations. They highlight that the ESOs, namely CEN, CENELEC and ETSI, will define how to comply with the EU AI Act's regulatory requirements for high-risk AI systems and that this delegation to ESOs raises concerns about accountability and legitimacy, as the organisations involved are private actors. This is particularly problematic considering that standards for high-risk AI systems are not purely technical but raise ethical considerations, and that in some organisations such as ETSI, where

97 OECD (2023), *OECD Guidelines for Multinational Enterprises on Responsible Business Conduct*, OECD Publishing, Paris, https://doi.org/10.1787/81f92357-en.
98 ISO (2012), ISO/IEC 17065:2012(en), *Conformity assessment — Requirements for bodies certifying products, processes and services*, available at: www.iso.org/obp/ui/en/#iso:std:iso-iec:17065:ed-1:v1:en (accessed 29 August 2024).
99 Cafaggi, F., A. Renda and R. Schmidt (2013), «Transnational private regulation», in *International Regulatory Co-operation: Case Studies, Vol. 3: Transnational Private Regulation and Water Management*, OECD Publishing, Paris, https://doi.org/10.1787/9789264200524-3-en.
100 ISEAL, *ISEAL Credibility Principles*, available at: www.isealalliance.org/defining-credible-practice/iseal-credibility-principles (accessed 28 August 2024).
101 Fairtrade International (2016), *Standard operating procedure: development of fairtrade standards: short summary*, available at: https://files.fairtrade.net/standards/SOP_short_Development_of_Fairtrade_Standards.pdf (accessed 29 August 2024) & FSC (2023), *Development and Revision of FSC Requirements Procedure (PRO) V(4-0),* available at: https://connect.fsc.org/document-centre/documents/resource/362 (accessed 29 August 2024).

direct industry participation is significant, the delegation creates risks of capture.[102] Other experts also explain that in ISO settings, where participation is organised by national delegations, there are no rules concerning the set-up of these delegations, which can consequently be dominated by industry actors.

Another difference lies in the existing instrument that is taken as the basis for developing assessment criteria. The criteria can be developed by the scheme owner (the case with most initiatives), be based on government regulation (for example, PEFC – the Programme for the Endorsement of Forest Certification),[103] or be aligned with international standards and guidelines (for example the OECD due diligence guidelines or UNESCO's ethical impact assessment tool[104]). This choice impacts not only the legitimacy of the standard and its enforceability, but also its ambition.

- ***Standard setting for public interest AI***

As has been mentioned, several internationally recognised principles and guidelines on artificial intelligence can be an inspiration for setting a standard on public interest AI. In addition, ongoing efforts to set standards can also be a benchmark. Building upon these works, would not only create coherence, but also provide legitimacy and inclusion into international efforts. It is however important to have a concrete and measurable standard with a high ambition to ensure that certified systems are public interest AI systems.

As recommended by the Forum, an independent and non-profit organisation should be responsible for standard setting. It should promote "active participation from stakeholders across the AI ecosystem, including developers, deployers, users, researchers, journalists and civil society organisations, with an emphasis on geographic, linguistic, cultural, and cognitive diversity. This would ensure that a wide range of diverse perspectives, including those of the Global Majority, inform the development and implementation of the certification system".[105] Even though working with CEN or ISO could provide some advantages in terms of legitimacy and standardised process, capture by AI industry actors is a risk. They have proven powerful in lobbying efforts regarding AI legislation,[106] Therefore, this paper recommends that the standard-setting process on public interest AI is led by non-governmental organisations and stakeholders.

3.2/ CERTIFICATION PROCESS

Once the standard is established, a certification process is needed to attribute the label. Such a system can include two steps: a self-assessment, and certification through an accredited independent body. These independent bodies may be accredited in a number of ways: either through an official

102 Cantero Gamito, M. & Marsden, C. (2024), Artificial intelligence co-regulation? The role of standards in the EU AI Act, *International Journal of Law and Information Technology*, 2024, 32.
103 Vermeulen, W.J.V et al (2010), *Roles of Governments in Multi-Actor Sustainable Supply Chain Governance Systems and the effectiveness of their interventions: An Exploratory Study*, Netherlands Environmental Assessment Agency (PBL), The Hague/Bilthoven, available at: www.pbl.nl/sites/default/files/downloads/PBL-2010-Other-roles-for-governments-needed-500411001.pdf (accessed 29 August 2024).
104 UNESCO (2023), *Ethical impact assessment: a tool of the Recommendation on the Ethics of Artificial Intelligence*, available at: https://unesdoc.unesco.org/ark:/48223/pf0000386276 (accessed 3 September 2024).
105 Forum on Information and Democracy (2024), *AI as a Public Good: Ensuring Democratic Control of AI in the Information Space*, available at: https://informationdemocracy.org/wp-content/uploads/2024/03/ID-AI-as-a-Public-Good-Feb-2024.pdf (accessed 28 August 2024).
106 Hartmann, T. (2023), AI Act : le gouvernement accusé d'avoir été influencé par un lobbyiste en situation de conflit d'intérêts, *Euractiv*, available at: www.euractiv.fr/section/intelligence-artificielle/news/ai-act-le-gouvernement-accuse-davoir-ete-influence-par-un-lobbyiste-en-situation-de-conflit-dinterets/ (accessed 29 August 2024).
Corporate Europe Observatory (2022), *Big Tech's last minute attempt to tame EU tech rules: Lobbying in times of trilogues*, available at: https://corporateeurope.org/en/2022/04/big-techs-last-minute-attempt-tame-eu-tech-rules (accessed 29 August 2024).
Perrigo, B. (2023), Exclusive: OpenAI Lobbied the E.U. to Water Down AI Regulation, *Time*, available at: https://time.com/6288245/openai-eu-lobbying-ai-act/ (accessed 29 August 2024).

accreditation process, or through recognition by other participants in the process (e.g., scheme owners). However, a detailed exposition of these processes is beyond the scope of this paper.

Both Fairly Trained and the Journalism Trust Initiative (JTI) include a self-assessment process. It is not just a more accessible process in terms of financial requirements, but is also a means of supporting the company or institution in improving its standards and procedures. The assessment system thereby works as a pedagogical and capacity-building tool. In both cases, the self-assessment process is complemented by an independent certification process.

Only once the company or institution has undergone the external certification process is it awarded the label. This process provides for more trust and legitimacy. Most certification mechanisms (i.e., FSC, JTI) choose to work with a range of accredited certification bodies, which are accredited according to ISO/IEC 17065 and therefore have to follow clear rules. Fairtrade has chosen to work with a single accredited certification body that they set up originally, namely Flocert. Generally, once an institution/product has received certification, the certification is valid only for a specific period of time and has to be renewed on a regular basis. This stipulation aims to ensure continuous compliance and potential evolutions and changes in the product.

- ***Certifying public interest AI systems***

Similarly, as with JTI, a two-step certification process could be envisioned for AI systems. The self-assessment component of the process would give AI system providers the opportunity to bring their practices up-to-standard.

However, the actual certification label should be provided only through an independent trusted certification assessment body, which has itself been certified according to ISO/IEC 17065. This is to ensure the integrity and trustworthiness of the certification. Such a procedure is of particular relevance for highly capable AI systems, whose inner workings can only be properly assessed by providing access to the system to an external auditor. This is also in line with the EU AI Act, which requires third-party conformity assessments for high risk AI systems (Article 6).

3.3/ MONITORING & ENFORCEMENT

Beyond standard setting, a certification mechanism generally includes activities for monitoring and enforcement of the standard, i.e., ensuring that the standard is respected by certified entities.[107] As identified by Marx & Depoorter, several mechanisms of institutional design can generate compliance by the certified entities.

Firstly, measures to ensure compliance include enforcement mechanisms such as audits (often conducted by certification bodies); bottom-up verification, where stakeholders can flag issues; transparency regarding audits, and consequences for non-compliance. Secondly, market incentives such as economic benefits and capacity building can also support compliance.[108] Finally, as discussed by Verbruggen, a certification mechanism can also refer a case to a public authority if the issue of non-compliance falls within the authority's competence.[109]

107 Cafaggi, F., A. Renda and R. Schmidt (2013), «Transnational private regulation», in *International Regulatory Co-operation: Case Studies, Vol. 3: Transnational Private Regulation and Water Management*, OECD Publishing, Paris, https://doi.org/10.1787/9789264200524-3-en.

108 Depoorter, C. and Marx, A. (2023), Fostering compliance with voluntary sustainability standards through institutional design: An analytic framework and empirical application, *Wiley*, available at : www.evidensia.eco/resources/2278/fostering-compliance-with-voluntary-sustainability-standards-through-institutional-design-an-analytic-framework-and-empirical-application/ (accessed 28 August 2024).

109 Verbruggen, P. (2013). Gorillas in the closet? Public and private actors in the enforcement of transnational private regulation. *Regulation & Governance*, 7, 512–532, available at: www.law.berkeley.edu/wp-content/uploads/2019/10/Verbruggen-Gorillas-2013.pdf (accessed 28 August 2024).

Most certification mechanisms studied for this paper include investigations and audits to identify fraud or non-respect. Fairtrade works with Flocert, which certifies and audits. Warning, reporting, or complaint mechanisms, such as Fairtrade's Accountability Watch or FSC's "report a problem" are also a common way to involve the public and interested stakeholders in monitoring and enforcement. In addition, ISEAL recommends that the scheme owner sets up a dispute resolution system to solve disagreements on complaints and compliance issues,[110] if they cannot be resolved directly with the certification body.

- ***Monitoring and enforcement of the public interest AI certification mechanism***

As recommended in the Forum's report, the certification mechanism should establish "an audit mechanism to ensure companies' ongoing compliance with the recommendations. Companies found to be non-compliant would lose this certification".[111] In addition to such an audit mechanism, and in line with the principles of stakeholder engagement and good practices of other certification mechanisms, it should also provide opportunities for stakeholders to flag issues and report non-compliance. In line with this, there is a need for a public repository enabling public auditing.[112] JTI, for example, has set up a repository of all certified media.[113]

A series of consequences for non-compliance can be envisioned starting with warnings for the company and increasing in severity up to withdrawal of the certification. As discussed below, capacity building and incentives schemes will further encourage compliance.

3.4/ FINANCING OF THE MECHANISM

A sustainable and independent financing model for the certification mechanism is necessary to ensure its relevance and trustworthiness. Most certification mechanisms are funded through multiple sources. These can include government funding or grants by private institutions; a certification fee paid by the company being certified (this is, however, often paid directly to the accredited certification body and not the scheme owner); and a fee paid by the company that uses the label. For example, retailers pay a fee to Fairtrade for every Fairtrade product they sell, or ICANN receives US$ 0.06 per domain registration, thereby creating a sustainable income.

- ***Financing the public interest AI certification mechanism***

Initially the certification mechanism could be set-up through government funding and other grants, while ensuring independence of the non-profit organisation managing the certification mechanism. However, to ensure sustainability and independence in the long run, a more reliable financing system would be needed. This could include a fee paid by every organisation being certified to the scheme owner, in addition to the fee that will be paid to the certification body. For companies over a certain threshold of turnover, this fee could be charged annually.

---

110 ISEAL (2023), *ISEAL Code of Good Practice for Sustainability Systems*, available at: www.isealalliance.org/sites/default/files/resource/2024-03/ISEAL-Code-of-Good-Practice-v1.0_ISEAL_final_public.pdf (accessed 28 August 2024).
111 Forum on Information and Democracy (2024), *AI as a Public Good: Ensuring Democratic Control of AI in the Information Space*, available at: https://informationdemocracy.org/wp-content/uploads/2024/03/ID-AI-as-a-Public-Good-Feb-2024.pdf (accessed 28 August 2024).
112 Forum on Information and Democracy (2024), *AI as a Public Good: Ensuring Democratic Control of AI in the Information Space*, available at: https://informationdemocracy.org/wp-content/uploads/2024/03/ID-AI-as-a-Public-Good-Feb-2024.pdf (accessed 28 August 2024).
113 JTI, *Community*, available at: www.journalismtrustinitiative.org/community (accessed 28 August 2024).

3.5/ GOVERNANCE OF THE MECHANISM

The governance of the certification mechanism, thus of the scheme owner as well as its decision-making body, is crucial as it defines the legitimacy of the certification mechanism, its risks to capture, and its ambition. As Cantero Gamito & Marsden highlight, the governance structure has an effect on the way that larger corporations can influence the mechanism. Therefore, the authors call for continuous stakeholder involvement in setting the standard, monitoring, enforcing, and refining it.[114] This is also in line with the ISEAL principles, which call for collaboration and stakeholder engagement.

Several questions therefore need to be answered: What kind of legal structure should manage the mechanism (i.e., be the scheme owner)? Where should it be based (i.e., in which jurisdiction)? And, which actors and with what decision-making power should be part of the governance structure? Cafaggi distinguishes different models. Certification mechanisms can be industry-driven, led by NGOs, led by experts or be multi-stakeholder.[115] Government-led certification, for example the Nutriscore and Dolphin label, also exists.[116]

Most certification mechanisms studied include NGOs, researchers, and representatives of marginalised groups, which is in line with the Forum's recommendations.[117] FSC also has a rule for 50% Global South representation. Some certification mechanisms also distinguish membership according to the actor, and assign different decision-making powers (i.e., voting rights) to different types of actors, as is the case for the Roundtable on Sustainable Palm Oil.[118] There are, however, two actors who raise challenges: namely industry actors – who, if they are present at the origin of the certification mechanism, turn it into an in-house scheme – and government actors, who could make the mechanism vulnerable to political considerations.

The next section discusses these two actors and their potential involvement:

- ***Government and public institutions***

Traditionally, governments had a limited role in their creation and governance in most certification mechanisms studied. This is the case for FSC, Fairtrade, and JTI (though not for those mechanisms created by governments, like Nutriscore or Blue Angel). No government representatives are on the governing boards of FSC, Fairtrade, and JTI. The Internet Corporation for Assigned Names and Numbers (ICANN) has developed a framework to involve governments through an advisory council, with representatives from more than 150 States. The Council has a liaison position on the board without voting rights. This council aims to ensure coherence between ICANN's decisions and government policies, thereby strengthening implementation.

However, Vermeulen et al. point out that governments have supported private certification mechanisms in implementing their work by providing economic incentives, tax benefits and public procurement policies, or promoting consumer awareness. Governments have also helped consumers in distinguishing

114 Cantero Gamito, M. & Marsden, C. (2024), Artificial intelligence co-regulation? The role of standards in the EU AI Act, *International Journal of Law and Information Technology*, 2024, 32.

115 Cafaggi, F. (2011), New Foundations of Transnational Private Regulation, *Journal of Law and Society*, vol. 38, no. 1, 2011, pp. 20–49, available at: www.law.berkeley.edu/wp-content/uploads/2019/10/Cafaggi-Foundations-of-TPR-2011.pdf (accessed 28 August 2024).

116 Marx, A. , Depoorter, C. , Lambert, S.R. and Fernandez de Cordoba, S. (2020), *Scaling up voluntary sustainability standards through sustainable public procurement and trade policy*, United Nations Forum on Sustainability Standards, available at: www.evidensia.eco/resources/999/scaling-up-voluntary-sustainability-standards-through-sustainable-public-procurement-and-trade-policy/ (accessed 29 August 2024).

117 Forum on Information and Democracy (2024), *AI as a Public Good: Ensuring Democratic Control of AI in the Information Space*, available at: https://informationdemocracy.org/wp-content/uploads/2024/03/ID-AI-as-a-Public-Good-Feb-2024.pdf (accessed 28 August 2024).

118 RSPO (2024), *Membership categories*, available at: https://rspo.org/as-an-organisation/membership/membership-categories/ (accessed 5 September 2024).

and identifying trustworthy labels.[119] Governments can also support the mechanism during its creation process through public endorsement or in providing funding. Such actions provide legitimacy. The lack of this endorsement is, according to representatives of some certification mechanisms, a challenge even today. Lack of legitimacy and recognition can be particularly difficult when competing labels are created.

- ***Industry and other actors being certified***

The inclusion of industry actors (in other words the organisations that are being certified) in the governance process is a divisive issue. Their exclusion creates the risk of their boycotting the mechanism, or of the mechanism not being suitable for application, while their inclusion creates a risk of industry capture, and of the lowest common denominator. Studies have shown that stakeholder input often includes an over-representation by industry groups, which weakens the standards.[120]

Certification mechanisms have opted for different approaches. FSC includes three chambers with equal power, an environmental one including mainly NGOs, a social one including trade unions, workers and indigenous people, and an economic one including industry actors. As the decision-making process is based on consensus this has led to difficulties in the past, but has also enabled having industry on board and fostering uptake. Fairtrade France chose a different model, providing industry with one seat out of 16 on the board, while the German and international branch excluded industry actors.

Different actors argue that industry needs to be included in one way or another to better implement the standard, take account of reality, and strengthen uptake. This also needs to be considered to prepare for and avoid a potential push-back by the sector. The AI sector is dominated by a few powerful companies that can have a significant impact on the success of the certification mechanism, therefore their inclusion needs to be carefully considered.

Industry buy-in could also be achieved through a rather symbolic participation. It should also be noted, particularly in the field of AI, that not all actors developing or deploying AI systems are private companies; they might also be non-profit, university, or governmental actors with different business models and objectives.

- ***Governance of the public interest AI certification mechanism***

The Forum's report therefore recommends "establishing an independent, non-profit organisation, or building upon an existing organisation to manage the certification process".[121] This kind of organisation should include representatives of users, researchers, journalists, and civil society organisations, and, building upon the model of the FSC, provide for equity between Global South and Global North members. It is also important to include representatives of the technically informed community, notably AI developers, which helps to ensure feasibility of the mechanism.

An advisory board without voting rights, as with ICANN, or a seat on the board without being in a position to block all decisions, could be attributed to a representative of States and to an AI system provider chosen as the representative by all certified providers. In addition, and in line with the ISEAL

119 Vermeulen, W.J.V et al (2010), *Roles of Governments in Multi-Actor Sustainable Supply Chain Governance Systems and the effectiveness of their interventions: An Exploratory Study*, Netherlands Environmental Assessment Agency (PBL), The Hague/Bilthoven, available at: www.pbl.nl/sites/default/files/downloads/PBL-2010-Other-roles-for-governments-needed-500411001.pdf (accessed 29 August 2024).

120 van der Ven, H., (2022), *Effects of stakeholder input on voluntary sustainability standards,* available at: www.evidensia.eco/resources/1322/effects-of-stakeholder-input-on-voluntary-sustainability-standards/ (accessed 29 August 2024).

121 Forum on Information and Democracy (2024), *AI as a Public Good: Ensuring Democratic Control of AI in the Information Space*, available at: https://informationdemocracy.org/wp-content/uploads/2024/03/ID-AI-as-a-Public-Good-Feb-2024.pdf (accessed 28 August 2024).

Code of Good Practice, the scheme owner should provide avenues for stakeholder participation beyond the governing board, paying particular attention to underrepresented stakeholders.[122]

3.6/ UPTAKE, SUSTAINABILITY, AND LEGITIMACY

A certification mechanism can only achieve impact if it is taken up by concerned actors, is sustainable over time, and resists attempts at abuse.

The ISEAL principles highlight that a certification mechanism should create value, which means "fairly rewarding the effort and resources that it takes for users to participate in the system".[123] Several measures can support value creation and facilitate uptake. According to Marx et al., these include consumer demand for the certified product, business demand, public procurement policies, and trade-based policies.[124] Links to accountability frameworks also encourage uptake, as certification is often considered to mean compliance with specific rules. Political endorsement and legal recognition are further means to strengthen uptake.

All these approaches are also important in creating legitimacy of the certification mechanism, as the mechanism does not derive its legitimacy from a state's sovereign authority. Rather, legitimacy comes through the market supply chain and follows three steps, as Bernstein and Cashore explain: early support, building widespread support that raises the costs for non-participation, and political legitimacy.[125] Creating legitimacy of the label is particularly important to prepare for potential counterattacks by the industry or other actors, and to prepare for the arrival of alternative labels, for example managed by the industry itself, i.e., in-house schemes.

Studies have shown that the emergence of fairtrade products and other sustainability labels creates consumer awareness, and fosters the uptake of such labels, as businesses want to meet consumer demands.[126] These labels also enable them to mitigate reputational risks and foster brand safety.[127]

Public procurement not only fosters uptake, but also provides legitimacy and credibility to a certification mechanism. Public procurement policies can also strengthen the requirements of a certification mechanism. Yet, these policies run the risk of discriminating against non-certified products, which respect similar standards without being officially certified.[128] In this respect, Article 43 of the European Union public procurement directive (2014/14) states that contracting authorities may require a specific label as proof for specific environmental, social or other characteristics. The article also sets out specific requirements for the label, such as objectively verifiable and non-discriminatory criteria, open and

122 ISEAL (2023), *ISEAL Code of Good Practice for Sustainability Systems*, available at: www.isealalliance.org/sites/default/files/resource/2024-03/ISEAL-Code-of-Good-Practice-v1.0_ISEAL_final_public.pdf.pdf (accessed 28 August 2024).

123 ISEAL, *ISEAL Credibility Principles*, available at: www.isealalliance.org/defining-credible-practice/iseal-credibility-principles (accessed 28 August 2024).

124 Marx, A. , Depoorter, C. , Lambert, S.R. and Fernandez de Cordoba, S. (2020), *Scaling up voluntary sustainability standards through sustainable public procurement and trade policy*, United Nations Forum on Sustainability Standards, available at: www.evidensia.eco/resources/999/scaling-up-voluntary-sustainability-standards-through-sustainable-public-procurement-and-trade-policy/ (accessed 29 August 2024).

125 Bernstein, S. & Cashore, B. (2007),*Can non-state global governance be legitimate? An analytical framework*, available at: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1748-5991.2007.00021.x (accessed 29 August 2024).

126 Redfern, A. and Snedker, P. (2002), *Creating Market Opportunities for Small Enterprises: Experiences of the Fair Trade Movement*, ILO, available at: www.ilo.org/publications/creating-market-opportunities-small-enterprises-experiences-fair-trade (accessed 28 August 2024).
Jodrell, D. and Kaoukji, D. (2022), *Exploring Fairtrade's impact: a review of research on Fairtrade from 2015-20,* Fairtrade International, available at: www.evidensia.eco/resources/1311/exploring-fairtrades-impact-a-review-of-research-on-fairtrade-from-2015-20/ (accessed 29 August 2024).
Nelson, V. (2009), *The Last Ten Years: A Comprehensive Review of the Literature on the Impact of Fairtrade,* available at: www.researchgate.net/publication/242598106_The_Last_Ten_Years_A_Comprehensive_Review_of_the_Literature_on_the_Impact_of_Fairtrade (accessed 29 August 2024).

127 Marx, A. , Depoorter, C. , Lambert, S.R. and Fernandez de Cordoba, S. (2020), *Scaling up voluntary sustainability standards through sustainable public procurement and trade policy*, United Nations Forum on Sustainability Standards, available at: www.evidensia.eco/resources/999/scaling-up-voluntary-sustainability-standards-through-sustainable-public-procurement-and-trade-policy/ (accessed 29 August 2024).

128 Marx, A. , Depoorter, C. , Lambert, S.R. and Fernandez de Cordoba, S. (2020), *Scaling up voluntary sustainability standards through sustainable public procurement and trade policy*, United Nations Forum on Sustainability Standards, available at: www.evidensia.eco/resources/999/scaling-up-voluntary-sustainability-standards-through-sustainable-public-procurement-and-trade-policy/ (accessed 29 August 2024).

transparent procedure, accessibility to all interested parties, and that "requirements are set by a third party over which the economic operator applying for the label cannot exercise a decisive influence".[129] Similarly, the French Digital Law encourages the use of open-source software for public administrations developing, buying, or using software (Article 16).[130]

Trade policies have also been used to promote sustainability labels, by integrating such labels in free trade agreements.[131]

Legal recognition of labels,[132] as well as political endorsement, can strengthen the label and its legitimacy. This can be seen in the case of JTI in the European Media Freedom Act. A mechanism on AI could be linked to existing processes, such as EU standardisation processes, OECD due diligence processes, or to UNESCO's *Ethical Impact Assessment*, which provides guidance to governments in assessing which AI system to privilege in public procurement.[133]

Finally the uptake, sustainability, and legitimacy of a certification mechanism depends on its integrity and resistance to abuse. The ISEAL principles highlight that a mechanism must respect the following principles: transparency about its owner, governance, and income; and impartiality, particularly when assessing performance, reliability and truthfulness.[134] The Code of Good Practice therefore suggests adopting an impartiality policy (on conflicts of interest); a diversity, equity, and inclusion policy; a safeguarding policy; a whistleblower protection policy, and a grievance policy.[135] As discussed above, the mechanism must avoid capture, particularly by government and industry actors. It must also avoid ethics-washing and be truthful to what it certifies. Therefore, it needs to protect its trademark through investigations, and by exacting consequences for infringement and non-compliance.

- ***Fostering uptake and legitimacy of a public interest AI certification mechanism***

As explained earlier, consumer demand can foster uptake of a label. Kieslich argues in this respect that public opinion plays an important role in fostering public interest AI.[136] However, there is currently little awareness regarding AI, or calls for responsible AI systems, by the public. Experiences of other actors trying to develop voluntary certification mechanisms for non-material goods, such as on safety, have struggled with uptake due to the lack of consumer demand. Other concepts, namely sustainability, have become more mainstream and could be built upon to promote consumer awareness of the importance of public interest AI. Therefore, as the Forum's report suggests, there is a need for a public awareness campaign.[137]

While, in the long run, the public interest AI label should be integrated into public procurement policies and trade policies, and potentially even legislation, in the short run support might be envisioned from

129 EU (2014), *Directive 2014/24/* https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32014L0024#d1e4888-65-1
130 Available at: www.legifrance.gouv.fr/jorf/article_jo/ JORFARTI000033203039 (Accessed: 9 February 2024).
131 OECD (2022), *The role of sustainability initiatives in mandatory due diligence: Background note on Regulatory Developments concerning Due Diligence for Responsible Business Conduct*, available at : https://mneguidelines.oecd.org/the-role-of-sustainability-initiatives-in-mandatory-due-diligence-note-for-policy-makers.pdf (accessed 28 August 2024).
132 OECD (2022), *The role of sustainability initiatives in mandatory due diligence: Background note on Regulatory Developments concerning Due Diligence for Responsible Business Conduct*, available at : https://mneguidelines.oecd.org/the-role-of-sustainability-initiatives-in-mandatory-due-diligence-note-for-policy-makers.pdf (accessed 28 August 2024).
133 UNESCO (2023), *Ethical impact assessment: a tool of the Recommendation on the Ethics of Artificial Intelligence,* available at: https://unesdoc.unesco.org/ark:/48223/pf0000386276 (accessed 3 September 2024).
134 ISEAL (2023), *ISEAL Code of Good Practice for Sustainability Systems*, available at: www.isealalliance.org/sites/default/files/resource/2024-03/ISEAL-Code-of-Good-Practice-v1.0_ISEAL_final_public.pdf.pdf (accessed 28 August 2024).
135 ISEAL (2023), *ISEAL Code of Good Practice for Sustainability Systems,* available at: www.isealalliance.org/sites/default/files/resource/2024-03/ISEAL-Code-of-Good-Practice-v1.0_ISEAL_final_public.pdf.pdf (accessed 28 August 2024).
136 Kieslich, K. (2024), *The role of public opinion on ethical AI principles and its implication for a common good-oriented implementation*, University of Hohenheim, available at: https://hohpublica.uni-hohenheim.de/server/api/core/bitstreams/27902e30-8ee2-4ce8-83fa-08f92a16db87/content (accessed 28 August 2024).
137 Forum on Information and Democracy (2024), *AI as a Public Good: Ensuring Democratic Control of AI in the Information Space*, available at: https://informationdemocracy.org/wp-content/uploads/2024/03/ID-AI-as-a-Public-Good-Feb-2024.pdf (accessed 28 August 2024).

political actors, such as the States of the Partnership for Information and Democracy, from international actors such as the G7, the UN, UNESCO, and the OECD, and from industry. The label could also be recognised as supporting the forthcoming OECD *Due Diligence Guidance for Responsible AI*, be aligned with UNESCO's *Ethical Impact Assessment*, and used as a monitoring tool for the implementation of the Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems. This political recognition is of particular importance in the AI context, which is characterised by a multitude of initiatives and would ensure making the voluntary certification mechanism competitive.

3.7/ MONITORING & LEARNING

Finally, any certification mechanism needs to be monitored and evaluated for continuous learning and adaptation. ISEAL suggests the principles of "measurable progress" and "continual improvement". As discussed above, this does not only include a process for revising the criteria, but it must also include processes for revising "intended outcomes, impacts, strategies, activities, and causal pathways"[138] and the governance structure. ICANN's governance structure, for example, is reviewed every three years to remain flexible and adapt to changes.

138 ISEAL (2023), *ISEAL Code of Good Practice for Sustainability Systems*, available at: www.isealalliance.org/sites/default/files/resource/2024-03/ISEAL-Code-of-Good-Practice-v1.0_ISEAL_final_public.pdf.pdf (accessed 28 August 2024).

CONCLUSION

This paper argues that an opportunity exists to create a voluntary certification mechanism for AI systems to encourage the development and use of public interest AI systems. This certification mechanism would respond to specific challenges and needs, such as tackling the information asymmetry between consumers and AI providers, creating an incentive scheme that is also available and useful in contexts where regulation is lacking or not desired, and putting the definition of what constitutes an ethical and responsible AI system in the hands of independent actors. The creation of such a mechanism is also timely, as it can be included in ongoing international efforts, including those led by the OECD and UNESCO, notably the OECD's forthcoming *Due Diligence Guidance for Responsible AI* and UNESCO's *Ethical Impact Assessment*. The voluntary certification mechanism for public interest AI could therefore constitute a trustworthy global institution to govern AI.

The paper further explores the design and specifications of the voluntary certification mechanism, and provides a preliminary suggestion of assessment criteria regarding the AI provider and the AI system, with a focus on criteria that can provide a proxy of the public interest character of the system. The paper also discusses the governance of the mechanism, raising important questions with regards to its sustainability, legitimacy, independence, and integrity. It suggests a governance framework inspired by the ISEAL Credibility Principles and argues that the mechanism should be owned and managed by an independent not-for-profit organisation with the inclusion of relevant actors, namely civil society, researchers, and affected communities. The paper further highlights the challenges of including industry actors, whose buy-in is crucial but whose capture can weaken the mechanism, as well as of government actors, who can support legitimacy but can also make the mechanism vulnerable to political considerations.

The paper does not pretend to provide the solutions to all questions, but rather aims to advance the debate on the establishment of a voluntary certification mechanism for AI systems. It provides the basis for more in-depth discussions among relevant actors, who will need to take decisions on assessment criteria and the governance structure. One way to take such decisions and define an approach more in detail would be by piloting a certification mechanism with interested AI providers.

BIBLIOGRAPHY

Bernstein, S. & Cashore, B. (2007),*Can non-state global governance be legitimate? An analytical framework*, available at: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1748-5991.2007.00021.x (accessed 29 August 2024).

Blue Angel (2020), *Resource and Energy-Efficient Software Products,* Basic Award Criteria Edition January 2020 Version 3, available at: https://produktinfo.blauer-engel.de/uploads/criteriafile/en/DE-UZ%20215-202001-en-Criteria-V3.pdf (accessed 28 August 2024).

Boot, E.R. (2024), The Public Interest: Clarifying a Legal Concept. *Ratio Juris*, 37: 110-129. https://doi.org/10.1111/raju.12401

Cafaggi, F., A. Renda and R. Schmidt (2013), "Transnational private regulation", in *International Regulatory Co-operation: Case Studies, Vol. 3: Transnational Private Regulation and Water Management*, OECD Publishing, Paris, https://doi.org/10.1787/9789264200524-3-en.

Cafaggi, F. (2011), New Foundations of Transnational Private Regulation, *Journal of Law and Society*, vol. 38, no. 1, 2011, pp. 20–49, available at: www.law.berkeley.edu/wp-content/uploads/2019/10/Cafaggi-Foundations-of-TPR-2011.pdf (accessed 28 August 2024).

Cantero Gamito, M. & Marsden, C. (2024), Artificial intelligence co-regulation? The role of standards in the EU AI Act, *International Journal of Law and Information Technology*, 2024, 32.

Casarosa, F. Cybersecurity certification of Artificial Intelligence: a missed opportunity to coordinate between the Artificial Intelligence Act and the Cybersecurity Act. *Int. Cybersecur. Law Rev.* 3, 115–130 (2022). https://doi.org/10.1365/s43439-021-00043-6

Collins, *Definition of 'fairness'*, available at: https://www.collinsdictionary.com/dictionary/english/fairness (accessed 28 August 2024).

Council of Europe (2024), *Council of Europe Framework Convention on Artificial Intelligence and human rights, democracy and the rule of law,* available at: https://rm.coe.int/1680afae3c

Corporate Europe Observatory (2022), *Big Tech's last minute attempt to tame EU tech rules: Lobbying in times of trilogues*, available at: https://corporateeurope.org/en/2022/04/big-techs-last-minute-attempt-tame-eu-tech-rules (accessed 29 August 2024).

Croteau, D. and Hoynes, W. (2006), *The business of media: corporate media and the public interest*, Pinge Forge Press

Department for Science, Innovation & Technology (2023), *Emerging Processes for Frontier AI Safety*, available at: https://assets.publishing.service.gov.uk/media/653aabbd80884d000df71bdc/emerging-processes-frontier-ai-safety.pdf (accessed 28 August 2024).

Depoorter, C. and Marx, A. (2023), Fostering compliance with voluntary sustainability standards through institutional design: An analytic framework and empirical application, *Wiley*, available at : www.evidensia.eco/resources/2278/fostering-compliance-with-voluntary-sustainability-standards-through-institutional-design-an-analytic-framework-and-empirical-application/ (accessed 28 August 2024).

European Commission (2024), *Data Governance Act explained*, available at: https://digital-strategy.ec.europa.eu/en/policies/data-governance-act-explained (accessed 4 September 2024).

EU (2024), *Artificial Intelligence Act,* available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689&qid=1722524514731

EU (2014), *Directive 2014/24*, available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32014L0024#d1e4888-65-1

Fairly Trained, available at: www.fairlytrained.org/ (accessed 28 August 2024).

Fairtrade International (2016), *Standard operating procedure: development of fairtrade standards: short summary*, available at: https://files.fairtrade.net/standards/SOP_short_Development_of_Fairtrade_Standards.pdf (accessed 29 August 2024).

Fairtrade International, available at: www.fairtrade.net (accessed 29 August 2024).

Fairtrade International, *Fairtrade Standards*, available at: www.fairtrade.net/standard (accessed 29 August 2024).

Forum on Information and Democracy (2024), *AI as a Public Good: Ensuring Democratic Control of AI in the Information Space*, available at: https://informationdemocracy.org/wp-content/uploads/2024/03/ID-AI-as-a-Public-Good-Feb-2024.pdf (accessed 28 August 2024).

Forest Stewardship Council, available at: https://fsc.org/en (accessed 29 August 2024).

FSC (2023), *Development and Revision of FSC Requirements Procedure (PRO) V(4-0)*, available at: https://connect.fsc.org/document-centre/documents/resource/362 (accessed 29 August 2024).

FSC, *What's in a label?* accessible at: https://fsc.org/en/what-the-fsc-labels-mean (accessed 29 August 2024).

Future of Life Institute, *Artificial Intelligence*, available at: https://futureoflife.org/cause-area/artificial-intelligence/ (accessed 28 August 2024).

Global Index on Responsible AI (2024), available at: www.global-index.ai/methodology (accessed 3 September 2024).

Hartmann, T. (2023), AI Act : le gouvernement accusé d'avoir été influencé par un lobbyiste en situation de conflit d'intérêts, *Euractiv,* available at: www.euractiv.fr/section/intelligence-artificielle/news/ai-act-le-gouvernement-accuse-davoir-ete-influence-par-un-lobbyiste-en-situation-de-conflit-dinterets/ (accessed 29 August 2024).

Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems, available at: www.mofa.go.jp/files/100573473.pdf

ICO, *What are the substantial public interest conditions?*, available at: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/special-category-data/what-are-the-substantial-public-interest-conditions/ (accessed 28 August 2024).

ISO,*Conformity assessment Certification*, available at : https://www.iso.org/certification.html (accessed 28 August 2024).

ISO (2012), *ISO/IEC 17065:2012(en), Conformity assessment* — Requirements for bodies certifying products, processes and services, available at: www.iso.org/obp/ui/en/#iso:std:iso-iec:17065:ed-1:v1:en (accessed 29 August 2024).

ISO/IEC 23894:2023, *Information technology — Artificial intelligence — Guidance on risk management*, available at: www.iso.org/standard/77304.html (accessed 28 August 2024).

ILO (2022), *ILO Declaration on Fundamental Principles and Rights at Work*, available at: www.ilo.org/ilo-declaration-fundamental-principles-and-rights-work (accessed 28 August 2024).

ISEAL, *ISEAL Credibility Principles*, available at: www.isealalliance.org/defining-credible-practice/iseal-credibility-principles (accessed 28 August 2024).

ISEAL (2023), *ISEAL Code of Good Practice for Sustainability Systems*, available at: www.isealalliance.org/sites/default/files/resource/2024-03/ISEAL-Code-of-Good-Practice-v1.0_ISEAL_final_public.pdf.pdf (accessed 28 August 2024).

Jodrell, D. and Kaoukji, D. (2022), *Exploring Fairtrade's impact: a review of research on Fairtrade from 2015-20*, Fairtrade International, available at: www.evidensia.eco/resources/1311/exploring-fairtrades-impact-a-review-of-research-on-fairtrade-from-2015-20/ (accessed 29 August 2024).

JTI, Journalism Trust Initiative, available at: www.journalismtrustinitiative.org (accessed 28 August 2024).

JTI, *Community,* available at: www.journalismtrustinitiative.org/community (accessed 28 August 2024).

JTI, CWA No. 17493, *CEN Workshop Agreement published on December 19th, 2019*, available at: www.jti-app.com/footer/cwa#standard_clauses (accessed 28 August 2024).

Kieslich, K. (2024), *The role of public opinion on ethical AI principles and its implication for a common good-oriented implementation*, University of Hohenheim, available at: https://hohpublica.uni-hohenheim.de/server/api/core/bitstreams/27902e30-8ee2-4ce8-83fa-08f92a16db87/content (accessed 28 August 2024).

Marx, A. , Depoorter, C. , Lambert, S.R. and Fernandez de Cordoba, S. (2020), *Scaling up voluntary sustainability standards through sustainable public procurement and trade policy*, United Nations Forum on Sustainability Standards, available at: www.evidensia.eco/resources/999/scaling-up-voluntary-sustainability-standards-through-sustainable-public-procurement-and-trade-policy/ (accessed 29 August 2024).

Max Havelaar Stiftung, *Qu'est ce que Fairtrade?*, available at: www.fairtrademaxhavelaar.ch/fr/quest-ce-que-fairtrade (accessed 28 August 2024).

Miller, K. (2023), *Introducing The Foundation Model Transparency Index*, Stanford University Human-Centered Artificial Intelligence, available at: https://hai.stanford.edu/news/introducing-foundation-model-transparency-index, (accessed 28 August 2024).

Ministère de l'économie, des finances, et de la souveraineté industrielle et numérique, *Georges Akerlof*, available at: www.economie.gouv.fr/facileco/georges-akerlof (accessed 28 August 2024).

Mission KI (2024), *Transparente Qualitäts- und Prüfstandards schaffen*, available at: https://mission-ki.de/ki-standards/ (accessed 4 September 2024).

Nelson, V. (2009), *The Last Ten Years: A Comprehensive Review of the Literature on the Impact of Fairtrade*, available at: www.researchgate.net/publication/242598106_The_Last_Ten_Years_A_Comprehensive_Review_of_the_Literature_on_the_Impact_of_Fairtrade (accessed 29 August 2024).

NIST Trustworthy and Responsible AI, *NIST AI 600-1, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*, available at : https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf (accessed 28 August 2024).

NIST (2024), *U.S. AI Safety Institute Signs Agreements Regarding AI Safety Research, Testing and Evaluation With Anthropic and OpenAI*, available at: www.nist.gov/news-events/news/2024/08/us-ai-safety-institute-signs-agreements-regarding-ai-safety-research (accessed 3 September 2024).

OECD (2024), *Recommendation of the Council on Artificial Intelligence*, available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (accessed 29 August 2024).

OECD (2023), *OECD Guidelines for Multinational Enterprises on Responsible Business Conduct*, OECD Publishing, Paris, https://doi.org/10.1787/81f92357-en.

OECD (2022), *The role of sustainability initiatives in mandatory due diligence: Background note on Regulatory Developments concerning Due Diligence for Responsible Business Conduct*, available at : https://mneguidelines.oecd.org/the-role-of-sustainability-initiatives-in-mandatory-due-diligence-note-for-policy-makers.pdf (accessed 28 August 2024).

Perrigo, B. (2023), Exclusive: OpenAI Lobbied the E.U. to Water Down AI Regulation, *Time*, available at: https://time.com/6288245/openai-eu-lobbying-ai-act/ (accessed 29 August 2024).

Ranking Digital Rights, available at: https://rankingdigitalrights.org/ (accessed 29 August 2024).

Redfern, A. and Snedker, P. (2002), *Creating Market Opportunities for Small Enterprises: Experiences of the Fair Trade Movement*, ILO, available at: www.ilo.org/publications/creating-market-opportunities-small-enterprises-experiences-fair-trade (accessed 28 August 2024).

Rowe, N. (2023), Millions of Workers Are Training AI Models for Pennies, *Wired*, available at: www.wired.com/story/millions-of-workers-are-training-ai-models-for-pennies/ (accessed 28 August 2024).

RSF (2023), *Paris Charter on AI and Journalism*, available at: https://rsf.org/sites/default/files/medias/file/2023/11/Paris%20Charter%20on%20AI%20and%20Journalism.pdf (accessed 28 August 2024).

RSPO (2024), *Membership categories*, available at: https://rspo.org/as-an-organisation/membership/membership-categories/ (accessed 5 September 2024).

Samuel, S. (2024),OpenAI insiders are demanding a “right to warn” the public, *Vox,* available at: www.vox.com/future-perfect/353933/openai-open-letter-safety-whistleblowers-right-to-warn (accessed 28 August 2024).

Thaler, R. H. and Sunstein, C. (2009), *Nudge: Improving Decisions About Health, Wealth, and Happiness*

UNESCO (2023), *Ethical impact assessment: a tool of the Recommendation on the Ethics of Artificial Intelligence*, available at: https://unesdoc.unesco.org/ark:/48223/pf0000386276 (accessed 3 September 2024).

UNESCO (2022), *Recommendation on the Ethics of Artificial Intelligence*, available at https://unesdoc.unesco.org/ark:/48223/pf0000381137 (accessed 28 August 2024).

van der Ven, H., (2022), *Effects of stakeholder input on voluntary sustainability standards,* available at: www.evidensia.eco/resources/1322/effects-of-stakeholder-input-on-voluntary-sustainability-standards/ (accessed 29 August 2024).

Vázquez, J-L., Lanero, A., García, J. A., Moraño, X (2023), *Segmentation of consumers based on awareness, attitudes and use of sustainability labels in the purchase of commonly used products*, available at: www.evidensia.eco/resources/2206/segmentation-of-consumers-based-on-awareness-attitudes-and-use-of-sustainability-labels-in-the-purchase-of-commonly-used-products/ (accessed 28 August 2024).

Verbruggen, P. (2013). Gorillas in the closet? Public and private actors in the enforcement of transnational private regulation. *Regulation & Governance,* 7, 512–532, available at: www.law.berkeley.edu/wp-content/uploads/2019/10/Verbruggen-Gorillas-2013.pdf (accessed 28 August 2024).

Vermeulen, W.J.V et al (2010), *Roles of Governments in Multi-Actor Sustainable Supply Chain Governance Systems and the effectiveness of their interventions: An Exploratory Study*, Netherlands Environmental Assessment Agency (PBL), The Hague/Bilthoven, available at: www.pbl.nl/sites/default/files/downloads/PBL-2010-Other-roles-for-governments-needed-500411001.pdf (accessed 29 August 2024).

WWF (2014), *The impact of Forest Stewardship Council Certification*, available at: https://wwfint.awsassets.panda.org/downloads/wwf_impact_forest_stewardship_council_certification_2014.pdf (accessed 28 August 2024).

Supported by

future of life INSTITUTE

Forum on Information and Democracy
CS 90247
75083 Paris Cedex 02, France
contact@informationdemocracy.org
https://informationdemocracy.org

Forum on Information & Democracy