

An Analysis of the New EU AI Act and A Proposed Standardization Framework for Machine Learning Fairness

Mike H.M. Teodorescu
Information School, RAISE
Center, IEEE Senior Member
University of Washington
Seattle, WA, USA
miketeod@uw.edu

Yongxu Sun
Information School, MSIM
Program
University of Washington
Seattle, WA, USA
yongxs@uw.edu

Haren N. Bhatia
Information School, MSIM
Program
University of Washington
Seattle, WA, USA
hnb24@uw.edu

Christos Makridis
Digital Economy Lab, Stanford
University, Stanford CA USA
University of Nicosia, Cyprus
makridis.c@unic.ac.cy

Abstract—The European Union's AI Act represents a crucial step towards regulating ethical and responsible AI systems. However, we find an absence of quantifiable fairness metrics and the ambiguity in terminology, particularly the interchangeable use of the keywords “transparency”, “explainability”, and “interpretability” in the new EU AI Act and no reference of transparency of ethical compliance. We argue that this ambiguity creates substantial liability risk that would deter investment. Fairness transparency is strategically important. We recommend a more tailored regulatory framework to enhance the new EU AI regulation. Furthermore, we propose a public system framework to assess the fairness and transparency of AI systems. Drawing from past work, we advocate for the standardization of industry best practices as a necessary addition to broad regulations to achieve the level of details required in industry, while preventing stifling innovation and investment in the AI sector. The proposals are exemplified with the case of ASR and speech synthesizers.

Keywords—AI transparency, explainability, ethical compliance, AI fairness, regulation, fairness criteria, auditing, strategic regulatory framework

I. INTRODUCTION

Artificial intelligence (AI) technologies have ushered us into a new era of innovation, with profound implications for society, economy, and governance. As AI systems become increasingly integrated into various aspects of daily life, the need for a regulatory framework that ensures these technologies are developed and deployed responsibly has become paramount. Already, a wide array of AI frameworks for considering the regulation of trustworthy use of AI have been introduced in the literature. The European Union (EU) has responded to this challenge with the EU AI Act (European Commission, 2021) [1], an expansive legislative effort aimed at promoting the ethical use of AI. This paper critically examines the EU AI Act, exploring its alignment with the principles of trustworthy AI and its potential impact on fairness and innovation, and proposes its blending with a set of industry standards that add more implementation detail to the EU AI Act's general frame and fairness transparency.

The EU AI Act is a comprehensive attempt to regulate AI across the EU, categorizing AI systems based on their potential

risk to individuals' rights and public safety. The Act delineates four levels of risk—minimal, limited, high, and unacceptable—with corresponding regulatory requirements for each category [1]. This risk-based approach is grounded in the EU's broader strategy on AI, which emphasizes the dual objectives of fostering technological excellence and ensuring trustworthiness (European Commission, 2021) [1].

Trustworthy AI encompasses a set of ethical principles and values that guide the development and deployment of AI systems. The EU's High-Level Expert Group on Artificial Intelligence has identified seven key requirements for trustworthy AI, including transparency, fairness, and accountability [2]. These principles are echoed in studies that advocate for AI systems to be designed in ways that respect human autonomy, prevent harm, and ensure justice and equity [3]; [4].

Fairness transparency of AI products is strategically important for legal compliance, liability avoidance, and public relations. However, this issue has not been thoroughly addressed in current regulations and business strategy literature.

II. PROPOSED SELF-REPORTING FRAMEWORK

Considering the legislature over broadly defined AI is not an optimal option for detailed definition and enforcement purposes, what could be the way forward? We suggest a hybrid system with a strong focus on standardization across the industry via a centralized public platform for self-assessed yearly audits addressing fairness, human augmentation, explainability, and interpretability, while still creating a strong legislature over high-risk predictive or generative services.

Several authors have argued for some form of framework for XAI, or for transparency or interpretability of AI [48], [50]; in addition, IEEE issued a “Guide for an Architectural Framework for Explainable Artificial Intelligence”. Our contribution is to align in a single framework AI transparency and fairness, such that fairness level can be derived in an explainable, possibly interpretable manner. In some respects, our framework has similarities with the study by Balasubramaniam et al. [52].

Partly following the literature [52], [47], we understand interpretability as the capacity of the AI to model a cause-effect relationship, where the model is readily understood by humans. Interpretability offers the clearest quantitative and qualitative clarifications to humans [47]. “Shallow” AI/ML, such as linear

predictors, are fully interpretable. Explainability refers to a translation by the machine of the established relationship between inputs (data) and outputs (predictions) in a way that is well understood by humans, although the details of establishing the connection are not available. Transparency means, in this study, the entirety of methods AI conveys meanings for the results to the human user; transparency covers interpretability and explainability, but vaguer, purely qualitative representations of the causality relations established between data and predictions also pertain to transparency.

A greater challenge is, however, to make transparent and explain how and why the predictions are fair and according to what fairness metric, out of the extant large number of metrics, they obey. In turn, fairness transparency allows auditability, which is key in algorithmic fairness.

To maintain fairness, transparency, interpretability, and explainability, we emphasize displaying the quantifiable fairness metrics, i.e., fairness through awareness, demographic parity, equalized opportunity, and equalized odds.

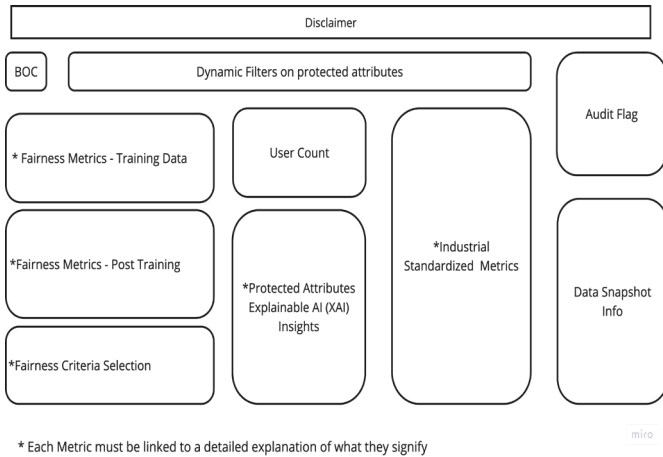


Fig. 1. AI Fairness & Transparency Dashboard Overview aimed to improve transparency of the fairness of the AI tools

Our proposed framework (see Figure 1) intends to function as a springboard for the standardization committees and the industry experts to build upon. The main elements and their motivation are briefly explained in Table 1. Fairness metrics must be applied to the data, whereas data affect what the AI system learns, and post training, for checking that the AI does not misinterpret parts of the data or distorts and bias the extracted information. The user oversees explicitly selecting the fairness criteria used, in agreement with the application. Industrial standardized metrics are more general, as they apply to all AI tools and may regard aspects less relevant to the developed tool. Auditing is mandatory to ensure that errors have not affected the development, that the AI tool is appropriate to the application and the task, and that the designers of the AI tools have not introduced involuntary biases in the data and the algorithm.

TABLE I. LEGEND FOR AI FAIRNESS AND TRANSPARENCY METRICS DASHBOARD

Label	Description
Disclaimer	Fairness metrics are relative; no AI is completely unbiased.
BOC (Bias Optimization Certificate)	Certifies trustworthiness from a standardizing body.
Dynamic Filters	View metrics based on protected attributes.
User Count	Number of users impacted by selected filters.
Industry Standard Metrics	Metrics for future AI applications, vendor suggested, and board approved.
Audit Flag	Flag to check if the company's reports have been audited by the standardizing board's auditing committee.
Data Snapshot Info	Includes timestamp, dataset names, and encrypted data for auditing.

In work from two of the authors, we employed Explainable AI (XAI) techniques, such as LIME, to identify biases within the decision-making process when opaque models are utilized by AI vendors [53]. Additionally, the standardization committee can attest to the display of custom-approved metrics if required to address the nuances of future AI technologies. We believe that the standardized framework be made with industry experts and provide baseline guidance on how it can be implemented. A flag variable could be displayed to know if the company has been audited independently and would link it to the historical findings of the previous audits. Lastly, a timestamp, dataset name, and encrypted data snapshot can be maintained for external auditing. Additionally, a blockchain-based solution for fairness checks may provide more confidence to users in terms of trusting the checks and the transparency of the entire process.

A. High-Risk Legislature and Standardization

Understanding the limitations in defining AI and future such definitions, we believe a method of standardization would be appropriate. These standardizations would take inspiration from other industries such as the International Standard Organization (ISO) for Quality Management Standards ISO9001, the International Electrotechnical Commission (IEC), and the ISO standard for Cybersecurity ISO/IEC 27032 [5]. We call upon such organizations to These organizations should have a unified and evolving approach to best practices, base security requirements, self-auditing, and explainability approaches. Services that follow these standards will be able to advertise a badge that would represent trustworthiness to their stakeholders. We posit that these standards should approach fairness from the perspective of organizational justice theory, addressing procedural and distributive fairness. Standardization would encourage flexibility thereby promoting innovation, would be developed speedily, and facilitate international consensus to a base level [6]; [5]. However, regulations will help provide enforcement to prevent the misuse of AI and its suppression of its unaware users. Therefore, we promote a hybrid approach of base-level standardization and enforceable legislature on high-risk Artificial intelligence, like the EU AI Act but with a more restrained approach to enforcement over a narrower audience of predictive

services, while still supporting innovation outside of the sand-box approach. To enforce High-Risk Artificial Intelligence, the legislature should consider the factors mentioned in Figure 1. Emboldening the importance of considering impact and having more transparency on the methodology used to pick ‘High-risk’ artificial intelligence would help enforcement, future proofing, and interpretability of their choices. Adding transparency to the fairness operation in the same framework would increase users’ confidence and attractiveness of the AI tools.

B. Fairness Metrics

The implementation of fairness criteria within the EU AI Act is notably ambiguous. The Act mentions the importance of avoiding discriminatory impacts and unfair biases but lacks specific metrics or methodologies for achieving fairness [1]. The act does not clarify this requirement with a strong definition. This gap is critical, given the complexity of operationalizing fairness in AI and the nuances of this ever-evolving field, as demonstrated by the discussion on fairness criteria such as fairness through unawareness, demographic parity, and equalized odds. Without clear guidance on fairness implementation, there is a risk that AI systems may perpetuate existing biases. We propose the following framework for quantified fairness in machine learning and artificial intelligence. Next, we sketch a framework for quantified fairness in ML / AI.

TABLE II. FAIRNESS CRITERIA DEFINITIONS

Fairness Metrics	Definitions
Fairness through Unawareness	This approach ignores the protected attributes in datasets. It hides the accountability of attributes for the predictor, which can be misleading [10].
Fairness through Awareness	Protected attributes are not ignored but acknowledged by ensuring similar individuals are classified similarly. However, this approach does not address group-level fairness.
Demographic Parity	Maintains the same positive prediction rate across groups.
Equalized Opportunity	Ensures the same true positive rate ($TP / (TP + FN)$) across different groups of protected attributes.
Equalized Odds	Ensures both true positive rate ($TP / (TP + FN)$) and false negative rate ($FN / (FN + TP)$) across different protected attributes for higher fairness.

The challenge in using automated decision-making systems or other forms of ‘AI’, as broadly defined in the EU AI act, comes down to the fairness of their models over individuals and groups, focusing on fairness across protected attributes. This paper will consider protected attributes as those defined by the U.S. Equal Opportunity Commission – race and color, religion, sex (including pregnancy, sexual orientation, or gender identity), national origin, age (40 or older), disability, and genetic information. While computational fairness is a difficult attribute to define, Morse employed organizational justice theory and addressed fairness as Distributive (fairness in outcome) and Procedural fairness (fairness in the process). To further delve into ensuring distributive and procedural fairness, there have been fairness criteria that can be quantified, namely, Fairness through

Unawareness, Fairness through Awareness, Demographic Parity, Equalized opportunity, and Equalized Odds. To quantify fairness, we employ fairness criteria to provide an empirical understanding of a model’s performance over protected attributes across individuals and groups.

C. Transparency, Explainability, and Interpretability

A significant number of the high-performing algorithms are subject to the black-box effect, which means that they are inherently non-interpretable to human understanding. Article 13(1) of the EU’s Artificial Intelligence Act states that “High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system’s output and use it appropriately.” Article (2) and (3) stressed that high-risk AI systems should be accompanied with clear instructions for use, which should also include “the characteristics, capabilities, and limitations for the performance of the high-risk AI systems”. However, neither EU AI Act nor other regulations have explicitly required transparency of the fairness of AI and of the process of embedding fairness in AI tools.

TABLE III. PROPOSED DEFINITIONS OF TRANSPARENCY, EXPLAINABILITY, AND INTERPRETABILITY TABLE TYPE STYLES

Terminology	Definition
Transparency	The accessibility and understandability of an AI system’s internal mechanisms and decision processes to humans without technical background.
Explainability	The ability of an AI system to provide human-understandable reasons for its outputs or decisions, at least by human experts in the domain of the respective reasons or decisions.
Interpretability	The extent to which humans can understand the outputs of an algorithm due to the complexity of the model’s design, when the humans are experts in the domain of the outputs.

Our paper centers on transparency, particularly emphasizing its two critical components: interpretability and explainability. These two essential aspects of transparency are frequently used interchangeably, despite the notable distinctions among these terms within the field of AI [47]. Even though the terms have yet to be given a clear universal definition, the law should start to establish explicit boundaries when regulations are to be imposed accordingly. To facilitate this, we propose definitions for each term to ensure a clearer distinction (Table 2).

D. Explainability through Explainable AI (XAI):

Benefits and Risks

The law does not explicitly mention Explainable Artificial Intelligence (XAI) when addressing the interpretability of high-risk AI systems. Instead, in Article 14, the law emphasizes that high-risk AI systems should be aware of the “available interpretation tools and methods.” Explainable AI (XAI) has developed as a crucial subfield to address the challenges associated with the limitations of AI models’ decision-making mechanisms. XAI offers post-hoc explainability techniques for models that are not inherently interpretable to humans, such as text

explanations, visual explanations, and local explanations. However, it is critical to acknowledge that many XAI tools currently available lack essential capabilities to effectively identify and address bias. This gap may contribute to a situation known as 'fairwashing,' where XAI tools incorrectly inform the researchers that a model is unbiased, while in truth, the biases persist undetected by the XAI tools. What is more, the connotation of transparency in the EU AI Act and XAI are different. The Act views transparency as a tool for promoting human rights and sustainable innovation, while XAI treats it more narrowly, aiming solely at unveiling algorithmic properties without considering its socio-technical implications, including fairness.

Another underlying problem that remains unsolved is the explainability gained from XAI approaches can hardly be quantified. The evaluations, comparisons and improvements of interpretability require quantifiable measures. Past literature overviews: on the measurement techniques developed for the goodness and accountability of the explanations offered by XAI (e.g., goodness checklist, explanation satisfaction scale, explanation trustworthiness) landed on the conclusion that more quantifiable and generalizable metrics should be developed. Therefore, if XAI is to be adopted, the law should avoid ambiguity by considering incorporating requirements for the establishment of standardized and quantifiable metrics. Similar challenges are also faced by interpretability, indicating a broad need for quantifiable metrics. To provide an example, Information Transfer Rate (ITR) measures the goodness of interpretability methods (e.g. LIME, COVAR) to intuitive human understanding. However, there are considerations to consider when developing such metrics. Firstly, different domains may necessitate distinct metrics tailored to the specific requirements of users within those domains. Secondly, while metrics are essential, they should be supplemented by qualitative human evaluations. After all, the ultimate objective is to foster understanding and trust in the model, including its fairness.

E. 2.5 Human Oversight in High-Risk Systems

In Article 14 (1), the high-risk AI systems were required to adopt a human oversight approach to enable the individuals to "correctly interpret the high-risk AI system's output". The incentive behind this approach is salient: past studies state that human-AI augmentation is the most promising path to fairness, and human oversight can promote societal values such as human dignity, legitimacy, and accountability. People are more likely to trust an AI system if they are informed that a human is involved in the decision-making process, aligning with the AIA's objective to foster public confidence to encourage the use of AI technologies to their full potential.

Despite the incentives behind it, adding a human into the loop is not the panacea to defuse AI risks (Agarwal et al., 2024). The Act itself has pointed out the danger of over-relying on the output of high-risk AI systems (automation bias); other limitations exist. The effectiveness of the human oversight approach relies on the assumption that humans can effectively oversee algorithmic decision-making. However, humans were found to perform undesirably on their overseeing tasks). This approach would also require the system to have adequate transparency so that the human overseers have substantive material to review and audit AI tools; other factors to be considered include the type of system and the level of training the overseers received.

III. DRAWBACKS AND EXISTING COUNTERMEASURES

If we consider the dynamic filtering criteria, we need to account for the number of scores that are reported. When considering all 7 legally protected attributes 5040 (7!) fairness metric scores need to be computed. Reporting at this granularity would be computationally intensive. Additionally, while we use random sampling of the data (Beutel et al., 2019), this approach is still costly because it excludes large portions of the data. Smaller subsets are more likely to result in a lower False Positive Rate (FPR). Moreover, determining an unbiased method of sampling remains a challenge. To address these issues, AI vendors can integrate adversarial learning and correlation regulations. The proposed solution integrates with libraries like the LinkedIn Fairness Toolkit (LiFT) to compute fairness in large-scale AI applications. LiFT is a reusable library that can be integrated with several API endpoints, by leveraging Apache Spark for general purpose cluster-computing to ensure data parallelism and fault tolerance. LiFT also provides an online service to measure model drift while monitoring fairness metrics in real-time. LiFT focuses on the following fairness metrics:

- Fairness in Training Data: KL and JS Divergences, L-p Norm and Total Variation Distances, Demographic Parity.
- Fairness in Metrics Post Model Training: Distance and Divergence metrics, Aggregate Fairness Metrics, and Differences in Model performance (across each sub-group using permutation testing).
- LiFT also provides pre-processing techniques, as well as custom metrics that may differ as per use case using User Defined Functions.

There are other ML fairness toolkits that we may consider, such as IBM's Fairness 360 tool, Microsoft AzureML Interpret, and Google's What-If, all of which work on their respective clouds or stand-alone technology and are not as flexible as LiFT.

We also recommend the reflection of explainability, and Interpretability metrics as mentioned in the previous section (e.g., Goodness checklist, explanation satisfaction scale, explanation trustworthiness), as well as the development of generalized quantifiable metrics, such as the Information Transfer Rate (ITR) to measure interpretability with extended covering of both the prediction and the fairness of it. Human-augmented feedback scoring could also be a future scope for developing new metrics in the field of explainability. All these scores should be auditable by an external body set up to track and enforce quality auditing of these self-attested scores.

IV. CASE STUDY: ASR SYSTEMS AND SPEECH SYNTHESIZERS

The domain of speech technology, comprising automated speech recognition (ASR) and speech synthesizers, is a narrow one in AI, although there are so numerous tools commercially available. As many of these tools can be found in homes for everyday use, many in the public and industry may consider this technology mature and beyond any ethical issues. While in most cases this may be true, numerous recent studies have revealed hidden ethical issues, such as limited coverage of the dialectal speech and of languages in less developed countries, discriminatory use of names and pronunciations, unsatisfactory

operation in the workplace, discriminatory operation for subjects with various medical conditions, elderly, and non-native speakers, and instances of enforced anthropological features.

This sub-field is probably too narrow to elicit a special law or even a federal regulation; instead, industry standards seem to be a good solution, but standards are lacking. Recent research has proposed multicriteria means to better characterize recognition systems in related applications (speaker recognition); similar means could be adopted for characterization of ASR systems. We suggest the development of standards for improving the ethical characterization of ASR systems and speech synthesizers based on a multi-parametric and multi-criteria evaluation that includes speakers and listeners using various dialects, from a minimal set of languages, and speakers/listeners with a large spectrum of medical conditions and ages, and under various levels of ambient noise. In addition, the standards should cover the level of diversity of using proper names and the direct and indirect use of anthropologic features.

V. CONCLUSION

We overviewed the European Union AI Act as a case study, with an emphasis on fairness testing, standardization and transparency. Since the new AI Act provides broader principles but leaves many implementation details still open, we proposed a standardization approach that presses upon a framework for the self-reporting of AI services with a high level of transparency of the fairness level and visibility to protect and trust the end users by providing them a peek into the ‘black box’ of their AI systems. The end users must be provided with direct links to the centralized public platform where the self-reported results would be published, just before the point of usage, enhancing the EU AI Act’s approach with a focus on interpretability and visibility at the point of usage. As a part of this standardized approach, all predictive services that deal with protected attributes, as previously defined, must self-audit at a pre-defined frequency, and report them on a centralized public platform.

While this will not be easily enforceable, complying companies will benefit strategically and may be given a badge similar to the CE marking, which will provide a signal of the quality of the software to users and other stakeholders. This, along with self-reported audits, would add a layer of trust between end users and enterprises.

On the legislative side, the residing commission can independently probe and audit these services and their reporting randomly. For quality self-reporting, the companies must log the metadata of the reported timestamps, and data snapshots, amongst other variables that we may consider depending upon the predictive services.

REFERENCES

- [1] Agarwal, A., Agarwal, H., & Agarwal, N. (2023). Fairness Score and process standardization: framework for fairness certification in artificial intelligence systems. *AI and Ethics*, 3(1), 267-279.
- [2] Agarwal, N., Moehring, A., Rajpurkar, P., & Salz, T. (2023). Combining human expertise with artificial intelligence: Experimental evidence from radiology (No. w31422). National Bureau of Economic Research.
- [3] Alikhademi, K., Richardson, B., Drobina, E., & Gilbert, J. E. (2021). Can explainable AI explain unfairness? A framework for evaluating explainable AI. *arXiv preprint arXiv:2106.07483*.
- [4] Aoki, N. (2021). The importance of the assurance that “humans are still in the decision loop” for public trust in artificial intelligence: Evidence from an online experiment. *Computers in Human Behavior*, 114, 106572.
- [5] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- [6] Beutel, A., Chen, J., Doshi, T., Qian, H., Woodruff, A., Luu, C., . . . Chi, E. H. (2019). Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements. *AEIS*.
- [7] Brennan-Marquez, K., & Henderson, S. E. (2019). Artificial intelligence and role-reversible judgment. *J. Crim. L. & Criminology*, 109, 137.
- [8] Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8), 832.
- [9] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness Through Awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*.
- [10] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*.
- [11] Enqvist, L. (2023). ‘Human oversight’ in the EU artificial intelligence act: what, when and by whom?. *Law, Innovation and Technology*, 15(2), 508-535.
- [12] Epley N, Waytz A, Akalis S, Cacioppo JT. When we need a human: Motivational determinants of anthropomorphism. *Social cognition*. 2008 Apr;26(2):143-55.
- [13] European Commission. (2021). Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts.
- [14] European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, 18: 3-32. <https://doi.org/10.1111/rego.12512>
- [15] Fraenkel, A. (2020). Fairness and Algorithmic Decision Making. MIT.
- [16] Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, 45, 105681.
- [17] Gyevar, B., Ferguson, N., & Schafer, B. (2023, October). Bridging the transparency gap: What can explainable AI learn from the AI Act?. In *26th European Conference on Artificial Intelligence* (pp. 964-971). IOS Press.
- [18] High-Level Expert Group on Artificial Intelligence. (2019). *Ethics Guidelines for Trustworthy AI*.
- [19] Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.
- [20] J. C. Wu and R. N. Kacker, "Statistical Analysis for Speaker Recognition Evaluation With Data Dependence and Three Score Distributions," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1-14, 2024, doi: 10.1109/TASLP.2023.3313447.
- [21] J.N. de Boer, A.E. Voppel, S.G. Brederoo et al., "Language disturbances in schizophrenia: the relation with antipsychotic medication", *NPJ Schizophr*, vol. 6, no. 24, 2020, [online] Available: <https://doi.org/10.1038/s41537-020-00114-3>.
- [22] Laux, J., et al. (2023). Fairness Criteria Implementation Chapter.
- [23] Laux, J., Wachter, S., & Mittelstadt, B. (2024). Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, 18(1), 3-32.
- [24] Laux, J., Wachter, S., & Mittelstadt, B. (2024). Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, 18(1), 3-32.
- [25] M.H.M. Teodorescu, L. Morse, Y. Awwad, G. Kane, Failures of Fairness in Automation Require a Deeper Understanding of Human–ML Augmentation. *MIS Quarterly* 45 (3), DOI: 10.25300/MISQ/2021/16535.

- [26] Mantello P, Ho MT, Nguyen MH, Vuong QH. Bosses without a heart: socio-demographic and cross-cultural determinants of attitude toward Emotional AI in the workplace. *AI & society*. 2023 Feb;38(1):97-119.
- [27] Mengesha Z, Heldreth C, Lahav M, Sublewski J, Tuennerman E. "I don't Think These Devices are Very Culturally Sensitive."—Impact of Automated Speech Recognition Errors on African Americans. *Frontiers in Artificial Intelligence*. 2021 Nov 26;4:725911.
- [28] Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4), 1-45.
- [29] Mori, T., & Uchihira, N. (2019). Balancing the trade-off between accuracy and interpretability in software defect prediction. *Empirical Software Engineering*, 24, 779-825.
- [30] Moriuchi E. An empirical study on anthropomorphism and engagement with disembodied AIs and consumers' re-use behavior. *Psychology & Marketing*. 2021 Jan;38(1):21-42.
- [31] Morse, L., Teodorescu, M. H. M., Awwad, Y., & Kane, G. C. (2022). Do the ends justify the means? Variation in the distributive and procedural fairness of machine learning algorithms. *Journal of Business Ethics*, Vol. 181, Iss. December 2022, pp. 1083–1095, <https://doi.org/10.1007/s10551-021-04939-5>.
- [32] Paats A, Alumäe T, Meister E, Fridolin I. Retrospective analysis of clinical performance of an Estonian speech recognition system for radiology: effects of different acoustic and language models. *Journal of Digital Imaging*. 2018 Oct;31(5):615-21.
- [33] Panigutti, C., Hamon, R., Hupont, I., Fernandez Llorca, D., Fano Yela, D., Junklewitz, H., ... & Gomez, E. (2023, June). The role of explainable AI in the context of the AI Act. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1139-1150).
- [34] Reitmaier, T., Wallington, E., Klejch, O., Markl, N., Lam-Yee-Mui, L.-M., Pearson, J., Jones, M., Bell, P., Robinson, S., Situating Automatic Speech Recognition Development within Communities of Under-heard Language Speakers. *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, April 2023, Article No.: 406, pp. 1–17, <https://doi.org/10.1145/3544548.3581385>.
- [35] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5), 206-215.
- [36] Schmidt, P., & Biessmann, F. (2019). Quantifying interpretability and trust in machine learning systems. *arXiv preprint arXiv:1901.08558*.
- [37] Schmidt, P., & Biessmann, F. (2019). Quantifying interpretability and trust in machine learning systems. *arXiv preprint arXiv:1901.08558*.
- [38] Shahin Shamsabadi, A., Yaghini, M., Dullerud, N., Wyllie, S., Aïvodji, U., Alaagib, A., ... & Papernot, N. (2022). Washing the unwashable: On the (im) possibility of fairwashing detection. *Advances in Neural Information Processing Systems*, 35, 14170-14182.
- [39] Smuha NA. Beyond the individual: governing AI's societal harm. *Internet Policy Review*. 2021 Sep;10(3).
- [40] Teodorescu, M. H. M., & Makridis, C. (2024). Fairness in Machine Learning: Regulation or Standards?, *Brookings Institution, Center on Regulation and Markets*, February 15 2024, <https://www.brookings.edu/articles/fairness-in-machine-learning-regulation-or-standards/>.
- [41] Teodorescu, M. H., & Yao, X. (2021). Machine Learning Fairness is Computationally Difficult and Algorithmically Unsatisfactorily Solved. 2021 IEEE High Performance Extreme Computing Conference (HPEC), Waltham, MA, USA, 2021, pp. 1-8, doi: 10.1109/HPEC49654.2021.9622861.
- [42] Teodorescu, M. H., Taraboanta, I., Andrian, S., Ghiorghe, A.C., Teodorescu, H. N., & Grigorie, M.. Fairness Constraints in Speech-to-Text Applications—Clinical Effects of Salivation. In *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, 2023, pp. 72-77. IEEE.
- [43] Teodorescu, MHM, Geiger, MK., Morse, L., (2024/4/22). Dangers of speech technology for workplace diversity. *Nature Machine Intelligence*, 6, 377–380 (2024). <https://doi.org/10.1038/s42256-024-00827-5>.
- [44] Vasudevan, S., & Kenthapadi, K. (2020). LiFT: A Scalable Framework for Measuring Fairness in ML Applications. *CIKM*.
- [45] Yeung, K. (2021). Regulation of High-Risk AI Applications – lost this one can you find it again?
- [46] Bussmann, N., Giudici, P., Marinelli, D. et al. Explainable Machine Learning in Credit Risk Management. *Comput Econ* 57, 203–216 (2021). <https://doi.org/10.1007/s10614-020-10042-0>.
- [47] Lukas-Valentin Herm, Kai Heinrich, Jonas Wanner, Christian Janiesch, Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability. *International Journal of Information Management* 69 (2023) 102538.
- [48] Hlongwane R, Ramabao K, Mongwe W (2024) A novel framework for enhancing transparency in credit scoring: Leveraging Shapley values for interpretable credit scorecards. *PLoS ONE* 19(8): e0308718. <https://doi.org/10.1371/journal.pone.0308718>
- [49] IEEE (2024) IEEE 2894-2024 Guide for an Architectural Framework for Explainable Artificial Intelligence, <https://standards.ieee.org/ieee/2894/11296/>.
- [50] Koen W. De Bock, Kristof Coussement, Arno De Caigny, Roman Słowiński, Bart Baesens, Robert N. Boute, Tsan-Ming Choi, Dursun Delen, Mathias Kraus, Stefan Lessmann, Sebastián Maldonado, David Martens, María Óskarsdóttir, Carla Vairetti, Wouter Verbeke, Richard Webe (2024), Explainable AI for Operational Research: A defining framework, methods, applications, and a research agenda. *European Journal of Operational Research*, Vol 317, 2, 249-272.
- [51] Winfield Alan F. T. , Booth Serena , Dennis Louise A. , Egawa Takashi , Hastie Helen , Jacobs Naomi , Muttram Roderick I. , Olszewska Joanna I. , Rajabiyazdi Fahimeh , Theodorou Andreas , Underwood Mark A. , Wortham Robert H. , Watson Eleanor, IEEE P7001: A Proposed Standard on Transparency. *Frontiers in Robotics and AI*. Vol. 8, 2021, URL=<https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2021.665729>. DOI=10.3389/frobt.2021.665729
- [52] Balasubramaniam, N., Kauppinen, M., Hiekkänen, K., & Kujala, S. (2022). Transparency and Explainability of AI Systems: Ethical Guidelines in Practice. In V. Gervasi, & A. Vogelsang (Eds.), *Requirements Engineering: Foundation for Software Quality - 28th International Working Conference, Proceedings* (pp. 3-18). *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Vol. 13216 LNCS). Springer. https://doi.org/10.1007/978-3-030-98464-9_1.