Ethical AI Governance: Methods for Evaluating Trustworthy AI

Louise McCormack[1,†], Malika Bendechache[1,†]

[1]ADAPT Research Centre, University of Galway, Ireland

Abstract

Trustworthy Artificial Intelligence (TAI) integrates ethics that align with human values, looking at their influence on AI behaviour and decision-making. Primarily dependent on self-assessment, TAI evaluation aims to ensure ethical standards and safety in AI development and usage. This paper reviews the current TAI evaluation methods in the literature and offers a classification, contributing to understanding self-assessment methods in this field.

Keywords

Artificial Intelligence (AI), Trustworthy AI (TAI), Evaluation Methods, AI Ethics, TAI Assessment

1. Introduction

Artificial intelligence (AI) is increasingly integrated into numerous sectors, making ethical considerations and trustworthiness in AI systems more critical than ever. Behavioural science is utilised to achieve objectives in areas such as climate change mitigation and educational attainment[1], a trend which also extends to Trustworthy AI (TAI). TAI is a crucial concept within the field of ethical AI, which encompasses the ethical considerations essential in the development and use of AI systems[2]. Leading TAI frameworks[3][4][5] incorporate behavioural science principles to ensure AI systems align with human values, considering their impact on behaviour and decision-making. Additionally, bidirectional human-AI alignment emphasises aligning AI to human values and enabling humans to adjust to AI advancements cognitively and behaviourally[6].

The European Commission Assessment List for Trustworthy AI (ALTAI)[7] and the European Union (EU) AI Act[8] are essential TAI guidelines, emphasising a human-centred, interdisciplinary approach. One recommended governance approach is establishing Standard-Setting Organisations that ensure minimum standards for testing, documentation and public reporting[9]. Despite the availability of various standards such as ISO/IEC 42001[10], evaluating and auditing AI systems remains challenging.

Several key surveys, such as those by Liu et al.[11] and Chamolaetal et al.[12], compile summaries of existing technical methods and technology in TAI. However, these surveys do not

AIEB 2024: Workshop on Implementing AI Ethics through a Behavioural Lens | co-located with ECAI 2024, Santiago de Compostela, Spain

[*] Corresponding author.

[†] These authors contributed equally.

louise.mccormack@adaptcentre.ie* (L. McCormack); malika.bendechache@adaptcentre.ie (M. Bendechache)

0009-0001-1782-1214 (L. McCormack); 0000-0003-0069-1860 (M. Bendechache)

focus on methods to score the areas of TAI. Ojewale et al.[13] propose a process for AI auditing, and although this work highlights the need for metrics and standards, it does not delve into the methods for calculating such metrics.

In this paper, we summarise and propose a classification and sub-classification for existing methods and systems to govern, evaluate, and score AI systems for trustworthiness aligned with the interdisciplinary human-centred approach taken by the EU. We also discuss challenges and future work in this area.

2. Methodology

2.1. Review Technique

Our survey was conducted through a Google Scholar query to identify methods used in the literature for TAI evaluation. In addition, we added articles, regulatory documentation, and ISO standards in this area through snowballing.

2.2. Research Questions

The following are the identified research questions for this review:

- Q1: What TAI evaluation methods and systems exist in the literature?
- Q2: What barriers to evaluating TAI are highlighted in the literature?

2.3. Research Search and Data Extraction Strategy

A search string for Google Scholar was designed to capture papers discussing topics in machine learning, trust and evaluation areas. Two researchers independently screened titles first and the abstract second to find papers that included TAI evaluation methods, resulting in 380 papers from the search string and an additional 12 papers through snowballing being reviewed. These papers were narrowed further, bringing the number of papers contributing to the core findings to 34. These papers were then summarised by both researchers and used to create a classification for the TAI evaluation methods.

3. Methods for Evaluating Trustworthy AI

In this section, we propose a classification for evaluating and scoring TAI. Of the papers reviewed, we found several approaches to AI scoring methods that considered various areas within TAI. Based on maturity and the type of solution proposed, we classed these papers into four categories: conceptual evaluation methods, Manual evaluation methods, Automated Evaluation Methods and Semi-Automated Evaluation Methods. In addition to this, we proposed a sub-classification based on the topic being evaluated. These sub-classifications are fairness & compliance evaluation, transparency evaluation, risk & accountability evaluation and trust & safety evaluation. As outlined in Figure 1, the most common approaches are conceptual approaches, indicating the lack of maturity in this field. This figure also shows the breakdown

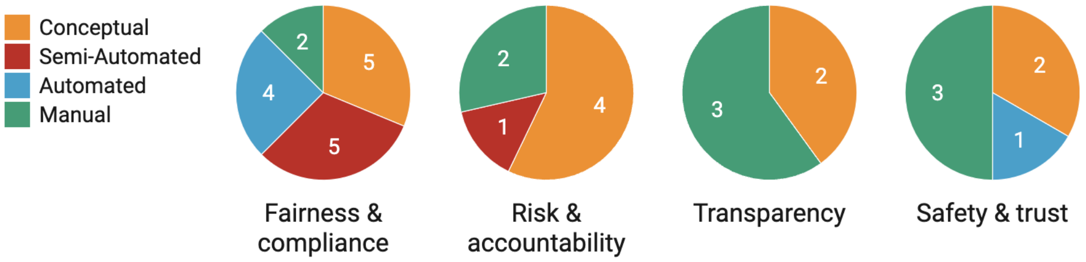

Figure 1: Comparison of evaluation methods by topic

of evaluation approaches by topic, particularly the number of automated and semi-automated evaluation methods already developed in fairness and compliance, one of the more researched areas of trustworthy AI.

3.1. Conceptual Evaluation Methods

The existing research includes several high-level governance frameworks that consider multiple dimensions of trustworthy AI throughout the AI lifecycle. Conceptual evaluation methods are high-level methods that do not provide implementation details or are not tested and validated. While conceptual frameworks in the literature can be holistic, they can also lack detail.

3.1.1. Fairness & Compliance Evaluation

Several conceptual approaches sought to evaluate and improve fairness and compliance in AI systems, introducing concepts like policy violation detection[14], using AI to define ethical behaviour[15][16] and automating fairness auditing[17][18]. Researchers used a variety of approaches in deciding what was fair, including incorporating existing established ethical guidelines[16], extracting ethical guidelines from social media[15], using a third-party regulator[18],[17] and extracting guidelines from policy documents[14].

3.1.2. Transparency Evaluation

Researchers proposed approaches that included evaluating transparency in areas such as healthcare[19] and finance[20]. The proposed framework by Lee[20] involved scoring fairness and interoperability, allowing humans to oversee and make conscious choices affecting both. The approach is context-conscious fairness and considers the trade-off between accuracy and interpretability and the trade-off between aggregate benefit and inequity. Trade-offs are benchmarked to make transparent, context-based, informed choices when using Machine Learning (ML) for decision-making. Jia et al.[19] proposed a framework to measure and improve technical robustness, safety, and transparency. It involved quantifying performance and XAI and establishing a trade-off between these trust properties for the ML algorithm selection for their healthcare use case.

3.1.3. Risk & Accountability Evaluation

Researchers also proposed conceptual governance frameworks that focused on risk management and accountability. These included ethical AI risk evaluation frameworks that built on the existing concepts such as operational design domain (ODD)[21][22]. The importance of defined safety boundaries was also highlighted[23][22].

Lu et al.[24] published a Responsible Artificial Intelligence (RAI) Pattern Catalogue, which was divided into multi-level governance patterns, trustworthy process patterns, and RAI-by-design product patterns, considering stakeholders at the industry, organisation, and team levels. This is important as researchers have shown engineers, legal experts, and users all require different levels of transparency from AI systems[25].

3.1.4. Trust & Safety Evaluation

Conceptual evaluation frameworks also addressed trust[26][27] and safety[28]. These frameworks firstly focused on identifying evaluation criteria or trust risk areas, then on methods to address these risk areas to improve trust[26][27].

Fisher et al.[28] discuss several use cases, focusing on safety-critical domains that require new standards and verification, validation, and certification methods. They include a classification of verification methods, including formal exhaustive static methods like model checking and theorem proving, non-exhaustive dynamic semi-formal methods like runtime verification and software testing, and non-exhaustive static methods like static analysis. The paper highlights the difficulty in certifying autonomous systems due to their complexity and evolving nature. Multiple stakeholder involvement creates complexity in establishing a consensus on acceptable ethical standards or evaluation criteria that do not disclose sensitive information.

Um et al.'s[26] layered trust framework includes a Trust Agent for data extraction, a Trust Analysis layer for computing trust metrics, and a Trust Management layer, addressing risk, fairness, security, design, traceability, data security, data privacy, and data pre-processing. Broderick et al.[27] created a taxonomy of trust in AI, which includes a process diagram for assessing the areas in which trust in ML can fail. They considered real-world use cases for finance, healthcare, and politics and subsequently provided ways to mitigate the risk and increase trust at each stage. Their conceptual process seeks to assess and mitigate the level of user trust, specifically the trust of an expert in their field at each stage.

3.2. Manual Evaluation Methods

One method proposed for assessing TAI is a manual questionnaire. Beyond the questions from the EU ALTAI[7] and ISO/IEC standards[29][10], six additional questionnaires were identified to score AI systems for trustworthiness. Manual questionnaires align with this area's regulation, considering multiple EU TAI principles. The disadvantage of the manual approach is that these questionnaires are typically time-consuming. This can lead to business constraints in completing the questions due to limited information about the external data the systems used[30].

3.2.1. Fairness & Compliance Evaluation

Approaches to improve fairness and achieve compliance in machine learning were proposed by researchers[31][32]. One approach was a practical questionnaire to help improve fairness by detecting bias[31]. A second approach to audit and score fairness in ML considered twelve metrics in this area[32]. The first six metrics focus on the stages of data collection, model development, feature selection and model performance—three metrics related to the human relationship with the model's decisions or predictions. The final metrics focus on assessing fairness from a broader social impact and include the three meta-components: cultural context, respect, and the research design process.

3.2.2. Transparency Evaluation

Transparency-focused questionnaires that focused on assessing the transparency of several TAI principles were also proposed by some researchers[33][30][34]. A notable questionnaire in the area of transparency is Bommasani et al.[34] who proposed The Foundation Model Transparency Index (FMTI), which included 100 indicators for transparency to be self-scored using a three-tier questionnaire and included benchmarks for leading organisations such as Open AI, AWS and Meta. Other researchers created separate transparency criteria for different tiers of stakeholders[33] and proposed using weighted questions using a 3-point scale for each question[30]. Transparency was also a consideration by researchers who looked at other areas such as user trust[35].

3.2.3. Risk & Accountability Evaluation

For security evaluation, researchers[36] scored existing questionnaire-based frameworks used in industry NIST[37], COBIT[38], ISO27001[29], and ISO42001[10] for their potential usage for AI's that incorporate Large Language Models (LLMs). Additionally, researchers developed a framework to evaluate the MITRE ATLAS[39] framework's effectiveness in protecting ML systems from poisoning attacks, scoring multiple TAI principles using a qualitative severity rating scale[40].

3.2.4. Trust & Safety Evaluation

Several questionnaire-based papers focused on trust and safety evaluation, typically asking users about their trust in various AI systems[41][42][35].

One approach was a simple unweighted user survey-based questionnaire, which scored several aspects of TAI evaluation, including intent and limitations, data, explainability, safety and robustness, audibility, and accountability[41]. Researchers also developed frameworks that used surveys to quantify and improve user trust by improving the transparency of the system[42][35]. Both papers successfully indicated a correlation between increased transparency and increased user trust in AI.

3.3. Automatic Evaluation Methods

This section includes papers investigating automated scoring methods for TAI Principles. Automatic methods ensure consistency in evaluation, however they rely on predefined metrics which do not exist for many aspects of trustworthy AI. The automated methods published to date are technical methods to evaluate and score the technical aspects of trustworthy AI with established methods and metrics.

3.3.1. Fairness & Compliance Evaluation

Several automated methods published to date are technical methods to evaluate and score fairness[43][44][45]. Notable methods include using data sampling techniques to measure and understand root causes of bias[44] and a sentence-based evaluation that used sentence likelihood difference (SLD) to calculate gender bias in LLMs[45]. Certification of fairness in AI systems was also considered by researchers who proposed a standard operating procedure (SOP) for fairness certification, Fairness Score and Bias Index, noting that different metrics would be needed to score pre-processing and in-processing and that the approach would be required to vary by use-case[46]. Researchers found that specific algorithms scored better for one set of individual features than others, indicating a link between fairness evaluation and algorithm selection[47].

3.3.2. Trust & Safety Evaluation

The automated evaluation of trust and safety of AI systems was also considered by researchers[48][43]. Researchers proposed an automated trust scoring process that used machine learning to develop a trust value for their use case of file sharing in peer-to-peer networks, automating a process to score the technical safety and likelihood of the file being dangerous[48]. Additionally, researchers developed a process that combined privacy and fairness evaluation, scoring both and proposing a trade-off for accuracy for each[43].

3.4. Semi-automated Evaluation Methods

This section covers approaches to scoring, which involve automated and manual steps. These methods are primarily in the area of fairness and compliance. They require a human at some stage, balancing automation and human efficiency. Researchers have shown the need to tailor evaluations by using case[49][18][50] and to incorporate considerations such as cultural differences in fairness evaluation[51]. In the case of healthcare, researchers reported that context was important in fairness evaluation for clinicians, noting a preference for a human-in-the-loop approach rather than a fully automated system[52].

3.4.1. Fairness & Compliance Evaluation

Researchers have proposed several semi-automated evaluation methods for fairness and compliance in AI[49][53][54][55][56]. A number of these frameworks were automated methods of fairness evaluation combined with a human element to set thresholds or decide trade-offs between metrics. One approach included developing transparent processes that mapped trade-offs between metrics[49], while a second involved injecting controls, wrapping existing operations and extending workflow primitives[53]. A third method included allowing a human to define the fairness requirement, specifying assumptions and assertions so that the tester can generate inputs that satisfy these assumptions and violate assertions[54]. A semi-automated user-centred approach to fairness evaluation called FairHIL (Fair Human-in-the-Loop) was developed that offers a visual user interface that provides a combination of visualisations including outcome features, feature intersection and causal graphs to help users identify bias and unfairness[55]. Users can add labels and adjust the feature weighting to retrain the model until they achieve an acceptable user fairness outcome. The tool focuses on accessibility and explainability for non-AI experts. Researchers also evaluated the effects of cultural differences in users interacting with the FairHil tool[56].

3.4.2. Risk & Accountability Evaluation

One paper proposed a semi-automated method for risk evaluation. This structured method provides an open vocabulary for AI risks (VAIR)[57], facilitating the automation of AI risk category identification, a required step for AI assessment in the EU AI Act.

4. Industry Tools for Evaluating TAI

In addition to the aforementioned academic works in evaluating TAI, various industry tools are in use today that aim to ensure AI systems adhere to ethical, legal, and performance standards. The most commonly used tools are manual questionnaire-based tools such as the ALTAI[7] and ISO/IEC 42001[10], which rely on self-assessments based on established principles, aligning with the self-assessment requirements of the EU AI Act[8]. These tools rely on human judgment and expert evaluations to identify risks and compliance issues, ensuring a thorough, albeit time-consuming, evaluation process. These manual methods are often supplemented by frameworks such as the NIST AI Risk Management Framework[37], which provides comprehensive guidelines for assessing safety, fairness, and transparency.

Automated assessment tools are becoming increasingly prevalent in the industry due to their efficiency and scalability. Tools like IBM's AI Fairness 360 and Microsoft Fairlearn are used to evaluate AI models for bias, fairness, and transparency without human intervention[58]. However, these are not accompanied by scientific, peer-reviewed papers evaluating their tools against the state-of-the-art works in this area[59]. Johnson et al.[59] publish an open-source toolkit called fair kit-learn, which is designed to support engineers in training fair machine learning models which found a better trade-off between fairness and accuracy than students using state-of-the-art tools sci-kit-learn and IBM AI Fairness 360[59].

These tools use sophisticated algorithms to identify and mitigate potential issues in AI systems, providing a scalable solution for large-scale AI deployments. Automated and semi-automated tools are particularly valuable, offering continuous monitoring and evaluation, enabling companies to maintain high standards of trustworthiness as AI systems evolve. Semi-automated tools such as Amazon SageMaker[60] combine automated algorithms with human oversight, ensuring a balance between efficiency and expert insight. Amazon SageMaker has features and tools that can be used to continuously monitor real-time data, concepts, bias, and feature attribution drives in models. These tools require human intervention at critical stages to set parameters and make interpretive decisions, ensuring that ethical and fairness considerations are adequately addressed.

Despite these advantages, recent research has highlighted several challenges practitioners face when using these tools. Practitioners find it difficult to translate real-world fairness concerns into quantifiable metrics that these toolkits can assess[61]. There is also a need for toolkits to be able to integrate more seamlessly into existing ML pipelines and to provide more guidance and resources for responsible usage[61]. Referring specifically to mitigating age bias in job selection using Microsoft Fairlearn and AI Fairness 360, researchers also found that significant human effort was required to make these toolkits work effectively to mitigate bias, making them impractical for usage in real-world applications[58].

5. Barriers to Trustworthy AI Evaluation

The complexity required for a complete evaluation of TAI presents several challenges. The barriers to evaluating TAI found in the literature include the following:

Diversity in Trustworthy AI Evaluation Method Evaluation methods exist for all aspects of TAI. However, the more mature areas of TAI have more advanced evaluation methods. For example, with several established methods, fully automated evaluation methods are available for fairness evaluation. Areas like risk and safety have some automatic and semi-automatic methods showing potential for more automation of technical aspects of AI where metrics are available. Evaluation approaches that considered less researched areas of TAI or holistic methods that considered multiple areas of TAI were primarily conception or manual methods.

Lack of Standardisation or Metrics for Evaluation Within the various TAI principles, there is a lack of consistency across all evaluation methods regarding what was being assessed. Even in similar industries using similar methods, the evaluation criteria or metrics used for evaluation were inconsistent. Regardless of the method used, this lack of consistency around evaluation criteria and metrics is a barrier to TAI evaluation and highlights a need to establish use case-specific benchmarks and acceptable thresholds for TAI evaluation.

Use Case Specific Evaluation Methods Required Clinicians found that context was essential when deciding acceptable evaluations for AI fairness. AI systems are complex, and their design varies by use case. Due to this complexity, the evaluation method will vary by use case. For example, evaluating a decision-making AI system requires a different approach versus other AI use cases such as an LLMs.

Human-in-the-Loop is Essential Although some automated methods exist to evaluate aspects of TAI, the semi-automated evaluation method is preferable if it integrates a human-in-the-loop. Additionally, due to a lack of maturity in many TAI principles, which have no metrics or automated methods for evaluation, a manual questionnaire-based stage is required for a comprehensive TAI evaluation. Additionally, even with more developed TAI principles such as fairness, a decision must be made manually to decide what is fair for the given use case.

Discrepancies Between Stakeholders Researchers found that different stakeholders all required different levels of transparency, meaning different methods and criteria for evaluation may be required for various groups of stakeholders. There are additional discrepancies between what stakeholders, such as AI and law experts, consider fair and what a layperson considers fair. There have been some semi-automated approaches to establishing ethical norms that can include multiple perspectives to combat this. One proposed conceptual method[15] involved extracting ethics from social media, which humans would then review for evaluation. Another approach was a semi-automated method[55] involving the development of a user interface with TAI metrics agreed upon by the AI developer that enabled human stakeholders to evaluate, make adjustments and decide trade-offs between TAI metrics.

Auditing and Third Party Accreditation is Required The research showed a need for governance in TAI evaluations that involved some form of access to the AI system. Several researchers who published conceptual governance frameworks proposed the inclusion of a third-party accreditation body which did this. These bodies would aim to provide the needed audits and governance for TAI evaluation. The research showed the potential to automate the audit and certification process for some TAI principles based on agreed metrics and benchmarks.

Fragmented Development and Accountability AI systems built using multiple organisations, including third-party data providers, face significant evaluation barriers. AI producers may lack access to necessary information from contributing organisations which they require for comprehensive TAI evaluations. For example, AI trained on data purchased from a third party might lack insight into data consent and acquisition processes, hindering thorough evaluation. In such instances, the AI producer struggles to assume accountability for development steps outsourced to other entities, making it challenging to perform a complete TAI assessment.

6. Future Directions for Trustworthy AI Evaluation

To successfully evaluate TAI, the literature calls for future AI systems to have ongoing semi-automated evaluation capabilities. Successful prototypes include using transparent or explainable models, with an interface allowing human decision-making of thresholds, trade-offs and/or definitions to be input into the model. This can be done by an expert in the field or a third-party accreditation body. Universal evaluation criteria and thresholds do not apply from one use case to the next, meaning that TAI principles would need a specific evaluation criterion for each use case.

There is a disconnect between the tools and research in this area. Tools used at the industry level have typically not been peer-reviewed and, when evaluated by researchers, are insufficient for comprehensive TAI evaluation versus the state of the art in the literature.

The findings of this paper have significant implications for AI policy. The research underscores the necessity for standardised evaluation frameworks to assess the trustworthiness of AI systems.

The current EU approach relies primarily on self-assessment and does not include methods or evaluation criteria for TAI evaluation, which the literature shows a clear need for. TAI standards developed by policymakers must be applied across use-case-specific AI applications to ensure ethical and fair practices. To facilitate comprehensive TAI evaluations for AI systems, governance frameworks in the literature propose third-party certification and standard methods and evaluation criteria, including metrics agreed upon by regulatory bodies based on their industry-specific needs and use cases. There is a disconnect between what policymakers, AI experts, and a standard non-expert user consider fair, along with differences based on culture, showing a need for more input from various laypeople to decide acceptable TAI evaluation approaches for individual use cases.

Authors and Affiliations All authors have reviewed and consented to the publication of the manuscript as presented. This research received partial support from the Science Foundation Ireland under grants 13/RC/2106P2 (ADAPT) and is co-funded by the European Regional Development Fund (ERDF).The data employed in this review are sourced from publicly available materials, including published research articles, ISO standards, books, and openly accessible databases and industry publications. All sources are duly cited and listed in the reference section of this paper. No new data was generated or collected specifically for this review.

References

[1] M. Hallsworth, A manifesto for applying behavioural science, Nature Human Behaviour 7 (2023) 310–322.

[2] D. Kaur, S. Uslu, K. J. Rittichier, A. Durresi, Trustworthy artificial intelligence: a review, ACM computing surveys (CSUR) 55 (2022) 1–38.

[3] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, et al., Ai4people—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations, Minds and machines 28 (2018) 689–707.

[4] S. Thiebes, S. Lins, A. Sunyaev, Trustworthy artificial intelligence, Electronic Markets 31 (2021) 447–464.

[5] D. Kaur, S. Uslu, A. Durresi, Requirements for trustworthy artificial intelligence–a review, in: Advances in Networked-Based Information Systems: The 23rd International Conference on Network-Based Information Systems (NBiS-2020) 23, Springer, 2021, pp. 105–115.

[6] H. Shen, T. Knearem, R. Ghosh, K. Alkiek, K. Krishna, Y. Liu, Z. Ma, S. Petridis, Y.-H. Peng, L. Qiwei, et al., Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions, arXiv preprint arXiv:2406.09264 (2024).

[7] A. HLEG, Assessment list for trustworthy artificial intelligence (altai) for self-assessment, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, 2020.

[8] E. Union, Final draft of the artificial intelligence act as of 2nd february 2024, https://artificialintelligenceact.eu/ai-act-explorer/, 2024.

[9] J. Laux, S. Wachter, B. Mittelstadt, Three pathways for standardisation and ethical disclosure by default under the european union artificial intelligence act. ssrn electron j, 2023.

[10] I. O. for Standardization, the International Electrotechnical Commission, Iso/iec 42001:2023, information technology – artificial intelligence – management system, Standard, 2023.

[11] H. Liu, Y. Wang, W. Fan, X. Liu, Y. Li, S. Jain, Y. Liu, A. Jain, J. Tang, Trustworthy ai: A computational perspective, ACM Transactions on Intelligent Systems and Technology 14 (2022) 1–59.

[12] V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, B. Sikdar, A review of trustworthy and explainable artificial intelligence (xai), IEEE Access (2023).

[13] V. Ojewale, R. Steed, B. Vecchione, A. Birhane, I. D. Raji, Towards ai accountability infrastructure: Gaps and opportunities in ai audit tooling, arXiv preprint arXiv:2402.17861 (2024).

[14] S. Shaikh, H. Vishwakarma, S. Mehta, K. R. Varshney, K. N. Ramamurthy, I. Wei, An end-to-end machine learning pipeline that ensures fairness policies, arXiv preprint arXiv:1710.06876 (2017).

[15] J. Buenfil, R. Arnold, B. Abruzzo, C. Korpela, Artificial intelligence ethics: Governance through social media, in: 2019 IEEE international symposium on technologies for homeland security (HST), IEEE, 2019, pp. 1–6.

[16] S. Umbrello, I. Van de Poel, Mapping value sensitive design onto ai for social good principles, AI and Ethics 1 (2021) 283–296.

[17] S. Park, S. Kim, Y.-s. Lim, Fairness audit of machine learning models with confi@inproceedingstial computing, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 3488–3499.

[18] D. van de Sande, J. van Bommel, E. Fung Fen Chung, D. Gommers, M. E. van Genderen, Algorithmic fairness audits in intensive care medicine: artificial intelligence for all?, Critical Care 26 (2022) 315.

[19] Y. Jia, J. McDermid, T. Lawton, I. Habli, The role of explainability in assuring safety of machine learning in healthcare, IEEE Transactions on Emerging Topics in Computing 10 (2022) 1746–1760.

[20] M. S. A. Lee, Context-conscious fairness in using machine learning to make decisions, AI Matters 5 (2019) 23–29.

[21] J. Roski, E. J. Maier, K. Vigilante, E. A. Kane, M. E. Matheny, Enhancing trust in ai through industry self-governance, Journal of the American Medical Informatics Association 28 (2021) 1582–1590.

[22] J. Mattioli, H. Sohier, A. Delaborde, K. Amokrane-Ferka, A. Awadid, Z. Chihani, S. Khalfaoui, G. Pedroza, An overview of key trustworthiness attributes and kpis for trusted ml-based systems engineering, AI and Ethics (2024) 1–11.

[23] G. Stettinger, P. Weissensteiner, S. Khastgir, Trustworthiness assurance assessment for high-risk ai-based systems, IEEE Access (2024).

[24] Q. Lu, L. Zhu, X. Xu, J. Whittle, D. Zowghi, A. Jacquet, Responsible ai pattern catalogue: A collection of best practices for ai governance and engineering, ACM Computing Surveys 56 (2024) 1–35.

[25] T. Van Nuenen, X. Ferrer, J. M. Such, M. Coté, Transparency for whom? assessing discriminatory artificial intelligence, Computer 53 (2020) 36–44.

[26] T.-W. Um, J. Kim, S. Lim, G. M. Lee, Trust management for artificial intelligence: A standardization perspective, Applied Sciences 12 (2022) 6022.

[27] T. Broderick, A. Gelman, R. Meager, A. L. Smith, T. Zheng, Toward a Taxonomy of Trust for Probabilistic Machine Learning, Science Advances 9 (2023) eabn3999.

[28] M. Fisher, V. Mascardi, K. Y. Rozier, B.-H. Schlingloff, M. Winikoff, N. Yorke-Smith, Towards a framework for certification of reliable autonomous systems, Autonomous Agents and Multi-Agent Systems 35 (2021) 1–65.

[29] I. O. for Standardization, Iso/iec 27001: Information technology - security techniques - information security management systems - requirements, Standard, 2013.

[30] J. Fehr, G. Jaramillo-Gutierrez, L. Oala, M. I. Gröschel, M. Bierwirth, P. Balachandran, A. Werneck-Leite, C. Lippert, Piloting a survey-based assessment of transparency and trustworthiness with three medical ai tools, in: Healthcare, volume 10, MDPI, 2022, p. 1923.

[31] J.-G. Lee, Y. Roh, H. Song, S. E. Whang, Machine learning robustness, fairness, and their convergence, in: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining, 2021, pp. 4046–4047.

[32] R. N. Landers, T. S. Behrend, Auditing the ai auditors: A framework for evaluating fairness and bias in high stakes ai predictive models., American Psychologist 78 (2023) 36.

[33] M. A. Chaudhry, M. Cukurova, R. Luckin, A transparency index framework for ai in education, in: International Conference on Artificial Intelligence in Education, Springer, 2022, pp. 195–198.

[34] R. Bommasani, K. Klyman, S. Longpre, S. Kapoor, N. Maslej, B. Xiong, D. Zhang, P. Liang, The foundation model transparency index, arXiv preprint arXiv:2310.12941 (2023).

[35] L. Guo, E. M. Daly, O. Alkan, M. Mattetti, O. Cornec, B. Knijnenburg, Building trust in interactive machine learning via user contributed interpretable rules, in: 27th International Conference on Intelligent User Interfaces, 2022, pp. 537–548.

[36] T. R. McIntosh, T. Susnjak, T. Liu, P. Watters, D. Xu, D. Liu, R. Nowrozy, M. N. Halgamuge, From cobit to iso 42001: Evaluating cybersecurity frameworks for opportunities, risks, and regulatory compliance in commercializing large language models, Computers & Security (2024) 103964.

[37] National Institute of Standards and Technology (NIST), Framework for Improving Critical Infrastructure Cybersecurity, Tech Report, National Institute of Standards and Technology (NIST), 2014. URL: https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf, https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf.

[38] Information Systems Audit and Control Association (ISACA), COBIT 2019 Framework: Governance and Management Objectives, ISACA, 2018. URL: https://www.isaca.org/bookstore/cobit/.

[39] MITRE Corporation, Mitre atlas, https://atlas.mitre.org/, 2023.

[40] C. Wymberry, H. Jahankhani, An approach to measure the effectiveness of the mitre atlas framework in safeguarding machine learning systems against data poisoning attack, in: Cybersecurity and Artificial Intelligence: Transformational Strategies and Disruptive Innovation, Springer, 2024, pp. 81–116.

[41] R. Dvorak, H. Liao, S. Schibel, B. Tribelhorn, Towards evaluating ethical accountability and trustworthiness in ai systems, Journal of Computing Sciences in Colleges 37 (2021)

11–22.

[42] J. Druce, M. Harradon, J. Tittle, Explainable artificial intelligence (xai) for increasing user trust in deep reinforcement learning driven autonomous systems, arXiv preprint arXiv:2106.03775 (2021).

[43] M. M. Khalili, X. Zhang, M. Abroshan, S. Sojoudi, Improving fairness and privacy in selection problems, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 8092–8100.

[44] J. Singh, A. Singh, A. Khan, A. Gupta, Developing a novel fair-loan-predictor through a multi-sensitive debiasing pipeline: Dualfair, arXiv preprint arXiv:2110.08944 (2021).

[45] K. G. BARZA, TOWARDS A ROBUST GENDER BIAS EVALUATION IN NLP, Ph.D. thesis, American University of Beirut, 2023.

[46] A. Agarwal, H. Agarwal, N. Agarwal, Fairness score and process standardization: framework for fairness certification in artificial intelligence systems, AI and Ethics 3 (2023) 267–279.

[47] M. Z. Alam, M. S. Rahman, M. S. Rahman, A random forest based predictor for medical data classification using feature ranking, Informatics in Medicine Unlocked 15 (2019) 100180.

[48] A. Alhussain, H. Kurdi, L. Altoaimy, A neural network-based trust management system for edge devices in peer-to-peer networks., Computers, Materials & Continua 59 (2019).

[49] M. S. A. Lee, L. Floridi, Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs, Minds and Machines 31 (2021) 165–191.

[50] S. Agarwal, Trade-offs between fairness, interpretability, and privacy in machine learning, Master's thesis, University of Waterloo, 2020.

[51] T. P. Pagano, R. B. Loureiro, F. V. Lisboa, R. M. Peixoto, G. A. Guimarães, G. O. Cruz, M. M. Araujo, L. L. Santos, M. A. Cruz, E. L. Oliveira, et al., Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and i@inproceedingstification and mitigation methods, Big data and cognitive computing 7 (2023) 15.

[52] S. Ryan, C. Nadal, G. Doherty, Integrating fairness in the software design process: An interview study with hci and ml experts, IEEE Access 11 (2023) 29296–29313.

[53] N. Antunes, L. Balby, F. Figueiredo, N. Lourenco, W. Meira, W. Santos, Fairness and transparency of machine learning for trustworthy cloud services, in: 2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W), IEEE, 2018, pp. 188–193.

[54] A. Sharma, H. Wehrheim, Automatic fairness testing of machine learning models, in: Testing Software and Systems: 32nd IFIP WG 6.1 International Conference, ICTSS 2020, Naples, Italy, December 9–11, 2020, Proceedings 32, Springer, 2020, pp. 255–271.

[55] Y. Nakao, L. Strappelli, S. Stumpf, A. Naseer, D. Regoli, G. D. Gamba, Towards responsible ai: A design space exploration of human-centered artificial intelligence user interfaces to investigate fairness, International Journal of Human–Computer Interaction 39 (2023) 1762–1788.

[56] S. Stumpf, E. Taka, Y. Nakao, L. Luo, R. Sonoda, T. Yokota, The need for user-centred assessment of ai fairness and correctness, in: Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, 2024, pp. 523–527.

[57] D. Golpayegani, H. J. Pandit, D. Lewis, To be high-risk, or not to be—semantic specifications and implications of the ai act's high-risk ai applications and harmonised standards, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 2023, pp. 905–915.

[58] C. Harris, Mitigating age biases in resume screening ai models, in: The International FLAIRS Conference Proceedings, volume 36, 2023.

[59] B. Johnson, J. Bartola, R. Angell, S. Witty, S. Giguere, Y. Brun, Fairkit, fairkit, on the wall, who's the fairest of them all? supporting fairness-related decision-making, EURO Journal on Decision Processes 11 (2023) 100031.

[60] D. Nigenda, Z. Karnin, M. B. Zafar, R. Ramesha, A. Tan, M. Donini, K. Kenthapadi, Amazon sagemaker model monitor: A system for real-time insights into deployed machine learning models, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 3671–3681.

[61] W. H. Deng, N. Yildirim, M. Chang, M. Eslami, K. Holstein, M. Madaio, Investigating practices and opportunities for cross-functional collaboration around ai fairness in industry practice, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 2023, pp. 705–716.