Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

FRICE: Enhancing AI Trust through a Robust Accountability System

Dr. Ravirajsinh Vaghela[1], Dr. Zalak Thakrar[2], Dr. Nirav Mehta[3], Ramji Goswami[4], Kaanchi Mukati[5]
[1]*School of Cyber Security and Digital Forensics, National Forensic Sciences University, Gandhinagar, India*
ravirajsing.vaghela@nfsu.ac.in
[2]*School of Open Learning, National Forensic Sciences University, Gandhinagar, India*
zalak.thakrar@nfsu.ac.in
(corresponding author)
[3]*Department of Computer Science, Shri V. J. Modha College of I.T - Porbandar, Porbandar, India*
mr.niravmehta1@gmail.com
[4]*School of Cyber Security and Digital Forensics, National Forensic Sciences University, Gandhinagar, India*
23.ramraja@gmail.com
[5]*School of Cyber Security and Digital Forensics, National Forensic Sciences University, Gandhinagar, India*
kaanchimukati@gmail.com

ARTICLE INFO

Received: 03 Sept 2024

Revised: 14 Oct 2024

Accepted: 27 Oct 2024

ABSTRACT

In the paper presented, a mathematical framework has been formulated on the basis of the FRICE Model to allow the evaluation of accountability in AI systems across five dimensions of assessment-Fairness, Robustness, Impact, Compliance and Explanatory Effectiveness-becoming necessary currently. The framework proposed here focuses on algorithmic assessment methods to ensure transparency, reliability and ethical governance in AI-based decision-making processes. Accountability score computation in the FRICE model employs a structured mathematical method, such as fairness-by Statistical Parity Difference (SPD) and Equal Opportunity Difference (EOD)-robustness by adversarial accuracy tests-impact by positive and negative outcome assessments-compliance and explanatory effectiveness. These features embed the principles of ethics into the system by ensuring compliance with legal stipulations and clarity in decision explanations. This framework takes weighted aggregation techniques incorporating parameters into account and translates them into a holistic accountability score, reflecting the performance of the system as a whole. Each parameter is presented together with its detailed formulas to ensure reproducibility and adaptation to various AI applications. Trade-offs will equally hinge on how ethical considerations are deeply integrated into the designs with fairness, inclusivity, and transparency to mitigate biases and ensure equitable results.

Keywords: accountability, fairness, robustness, impact, compliance, explanatory tools

INTRODUCTION

Artificial intelligence (AI) systems are increasingly relied upon in high-stakes decision-making, raising the need for accountability. AI accountability ensures that systems reliably align with societal and ethical standards while minimizing harm. AI accountability plays a critical role in guiding the ethical, reliable, and sustainable development and deployment of AI systems across various global domains. With AI systems becoming pervasive in industries such as healthcare, finance, education, and government, ensuring accountability directly addresses key societal, legal, and ethical concerns while aligning with global AI trends.

A comprehensive body of work exists identifying methodologies for AI accountability across traceability mechanisms, failure analysis, and external oversight, but most contributions focus on conceptual frameworks ([1, 6, 4]) and tool proposals ([3, 8, 18]) rather than validated, scalable real-world implementations, with critical gaps in standardization, integration, and practical failure analysis approaches ([5, 10, 4]).

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

KEY REASONS WHY AI ACCOUNTABILITY IS CRUCIAL IN GLOBAL TRENDS

A. Building trust and encouraging responsible AI adoption

Accountability fosters trust by promoting transparency, interpretability, and robust performance of AI systems. Users and stakeholders are more likely to adopt AI solutions when they understand how decisions are made, and safeguards are in place to address adverse outcomes [1, 3]. Frameworks like CERTIFAI, which provide fairness, robustness, and transparency evaluations, are examples of accountability mechanisms designed to improve stakeholder confidence in AI [1, 2].

B. Addressing ethical concerns, bias and discrimination

Accountability mechanisms mitigate systemic bias and unfair outcomes by ensuring fairness and equity in AI decision-making. This is particularly critical in high-stakes applications such as hiring, credit scoring, and criminal justice, where biased models can create discriminatory outcomes [8, 11]. Tools like causal fairness frameworks (e.g., Structural Causal Models) provide formal methodologies to detect and address discrimination while adhering to ethical and societal norms [8, 14, 19].

C. Ensuring legal compliance across jurisdictions

Global AI regulations, such as the European Union's AI Act, the General Data Protection Regulation (GDPR), and emerging U.S. algorithmic accountability laws, necessitate the alignment of AI systems with legal requirements, including transparency, fairness, and explainability [18, 13]. Accountability frameworks that incorporate legal compliance, such as by evaluating fairness under disparate impact scenarios, are necessary for AI systems operating in multiple regulatory environments [14, 19].

REVIEW OF LITERATURE

A major leap in this domain has been the development of model-agnostic frameworks that assess multiple accountability principles in a unified manner:

1. CERTIFAI ([1, 2]): Combines fairness assessment, robustness evaluation (via CERScore), and counterfactual reasoning for explainability into an integrated auditing framework. CERTIFAI is applicable to black-box models and supports fairness-sensitive applications.
2. ComplAI ([3, 4]): Introduces a "Trust Factor," a composite metric evaluating fairness, robustness, explainability, and susceptibility to drift. It facilitates comparison and improvement of accountability across diverse machine learning models.
3. AVOIR ([18]): Focuses on fairness auditing in real-time by monitoring runtime violations of fairness metrics, connecting accountability with adaptive system evaluation.

These frameworks signify a shift from the siloed exploration of fairness, robustness, and explanation toward comprehensive accountability metrics. However, none fully address all five principles (e.g., legal compliance or societal impact).

State of the Art: Emerging Techniques Fairness and Explainability Synergies Recent developments target hybrid approaches to integrate multiple principles:

Fairness-aware explanations ([6, 7, 11]): Novel methods like fairness explanation metrics or optimal transport apply fairness constraints while producing transparent explanations.

Causal Fairness Models ([8, 14]): Structural causal models provide fairness-aware explanations and compliance diagnostics, aligning accountability with real-world norms.

Year	Contribution to Accountability Models
2018	**Datasheets for Datasets (Gebru et al.)**: Introduced dataset-level traceability for transparency [**15**].
2019	**FactSheets (IBM Research)**: Proposed complete documentation for AI systems' purpose, performance, provenance, and adjustments [**11**].

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

2020	**SMACTR Framework**: Lifecycle audit model focusing on traceability and ethical foresight [6].
2020	**GAF (Global-view Accountability Framework)**: Framework targeting liability for automotive systems and insurance [1].
2021	**Traceability Reviews (Kroll, W3C PROV Extensions)**: Systematic review of metadata standards and challenges [9, 10].
2021	**Algorithm Audit Framework (Brown et al.)**: Stakeholder-driven ethical audits for governance [16].
2022	**Knowledge Graphs for Accountability (Naja et al.)**: Expanded traceability scope using semantic structures [8].
2023	**LLM-Specific Auditing Frameworks (Floridi)**: Introduced three-layered audits for large language models [4].
2024	**Criterion Audit Framework (Lam et al.)**: Structured auditing approaches to govern decisions in hiring systems [2].
2024	**Continuous Auditing Model (AuditMAI)**: Framework for real-time auditing infrastructure across the AI lifecycle [18].
2018	**Datasheets for Datasets (Gebru et al.)**: Introduced dataset-level traceability for transparency [15].
2019	**FactSheets (IBM Research)**: Proposed complete documentation for AI systems' purpose, performance, provenance, and adjustments [11].

Table 1. Trend: Growing focus on regulations and standardization

PROPOSED FRAMEWORK

Accountability Score Calculator

Proposed FRICE formula to compute an overall accountability score for AI model.

Key Components of the FRICE Formula

1. Fairness: Measures the equitable treatment of different groups.
2. Robustness: Assesses how well the model performs under adversarial or noisy conditions.
3. Impact: Captures the societal and ethical effects of deploying the AI system.
4. Compliance: Ensures adherence to regulatory and ethical guidelines.
5. Explanatory Effectiveness: Evaluates the quality and coverage of explanations provided by the system.

Each parameter is weighted to reflect its relative importance in the overall accountability score. The weights are configurable and must sum to 1.

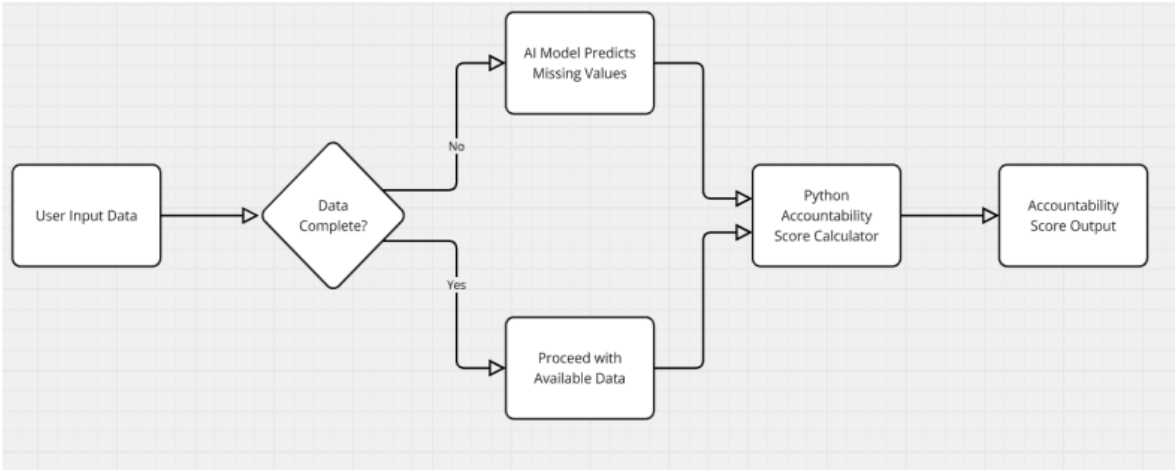

Fig. 1. System Architecture

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Defining the Parameters:

The presented flowchart demonstrates a hybrid system architecture designed for the calculation of accountability scores within an AI accountability framework. The system integrates two primary processes: a rule-based Algorithm and an AI model for prediction of missing values. The flowchart follows these structured steps:

1. User Input Data: The process begins with user-provided input data, which may include partial or complete information about the key parameters necessary for accountability score computation.
2. Data Completeness Check: A decision node verifies the completeness of the input data:

   Yes Path: If the data is complete, the system proceeds directly to the next module.

   No Path: If the data is incomplete, the AI model is invoked to predict the missing parameters.
3. AI Model Predicts Missing Values: The AI model uses a trained regression approach to impute missing parameter values based on learned relationships from historical data. The predicted values are fed into the next module for accountability score computation.
4. Algorithm Accountability Score Calculator: The Python-based accountability calculator takes the available and predicted parameters as inputs and calculates the accountability score using the FRICE formula, which incorporates Fairness (F), Robustness (R), Impact (I), Compliance (C), and Explanatory Effectiveness (E).
5. Accountability Score Output:The final accountability score is generated and presented as the output. This score reflects the holistic evaluation of the system's fairness, robustness, impact, compliance, and explainability.

D. Algorithms for FRICE Formula:

Algorithm Calculate_FRICE_Score:

Input: F, R, I, C, E // Fairness, Robustness, Impact, Compliance, Explanatory Effectiveness

Input: W_F, W_R, W_I, W_C, W_E // Weights for FRICE parameters

Output: A // Accountability Score

Step 1: Initialize weights such that W_F + W_R + W_I + W_C + W_E = 1

Step 2: Check if F, R, I, C, E ∈ [0, 1]

If any parameter is not in range [0, 1], return “Invalid Input”

Step 3: Compute weighted contributions:

Fairness_Contribution← W_F * F

Robustness_Contribution← W_R * R

Impact_Contribution← W_I * I

Compliance_Contribution← W_C * C

Explanation_Contribution← W_E * E

Step 4: Calculate Accountability Score:

A←Fairness_Contribution + Robustness_Contribution+ Impact_Contribution + Compliance_Contribution+ Explanation_Contribution

Step 5: Return A

End Algorithm

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

E. Defining the Parameters

1) Fairness[2]:

Fairness[2] (F) in the context of AI accountability ensuresthat the model treats different groups equitably in terms of outcomes and opportunities. It is calculated using two key metrics: Statistical Parity Difference (SPD)[2] and Equal Opportunity Difference (EOD)[2].

Metrics:

a) Statistical Parity Difference (SPD)[2] :

Measures the difference in positive outcome rates between two groups (G=a and G=b)

Formula: SPD=|P(Y=1|G=a)−P(Y=1|G=b)|

Examples:

- Group A (men) and Group B (women) apply for loans.
- P(Y=1|G=a): Proportion of men approved for loans.
- P(Y=1|G=b): Proportion of women approved for loans.
- SPD = 0: Perfect fairness; equal approval rates.
- SPD = 0.2: Significant disparity; unfair.

b) Equal Opportunity Difference (EOD) [2]:

Measures the difference in true positive rates (TPR) across groups.

Formula: EOD=|TPR(G=a)−TPR(G=b)|

TPR=True Positives+(False NegativesTrue Positives)

Examples:

- Group A (men) and Group B (women) apply for jobs.
- TPR(G=a): Proportion of qualified men hired.
- TPR (G=b): Proportion of qualified women hired.
- EOD =0: Perfect fairness; equal hiring rates.
- EOD = 0.15: Significant disparity; unfair.

Normalization:

To normalize F and ensure it lies between 0 and 1:

- F=1−max(SPD,EOD)
- F=1: Perfect fairness.
- F=0 Maximum unfairness.

2) Robustness[3]:

Robustness [3](R) is a critical parameter to assess how well an AI model performs under varying conditions, such as adversarial attacks, noisy inputs, or small perturbations in the input data. It ensures that the model's predictions remain reliable and consistent even in non-ideal scenarios.

Metrics:

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

a) Baseline Accuracy (BA) [3]

Measures the fraction of correct predictions on clean (non-adversarial) data.

Formula : BA=Total Predictions/Correct Predictions

b) Adversarial Accuracy (AA)[3]

Measures the fraction of correct predictions when the model is subjected to adversarial perturbations.

Formula: AA=Total Adversarial Predictions/Correct Adversarial Predictions

c) Combined Robustness Score (R) [3]

Combines baseline accuracy and adversarial accuracy to provide a unified measure of robustness.

Formula: R=(BA+AA) /2

Normalization of Robustness:

To ensure consistency and comparability, the robustness score can be normalized using Min-Max scaling:

Rn=(R−Rmin) (Rmax−Rmin)

3) Impact[4]

Impact[4] (I) evaluates the societal and ethical consequences of deploying an AI model. It captures both positive contributions (e.g., enhancing user experience, societal benefit) and negative effects (e.g., harm, bias, or risks). The overall impact is a normalized score ranging from 0 to 1.

Terms in Impact Calculation:

a) Positive Impact Score (PI) [4]:

Measures the proportion of positive effects produced by the model.

Formula: PI=Positive Effect Count/Total Effect Count

b) Negative Impact Score (NI)[4]:

Measures the proportion of negative effects produced by the model.

Formula: NI=Negative Effect Count/Total Effect Count

c) Maximum Possible Impact (Max Impact)[4]:

A constant representing the maximum theoretical impact of the system.

Default is often set to 1 for normalization.

d) Impact Score (I):

Combines PI[4] and NI[4] into a single normalized score.

Formula: I=(PI−NI)/Max Impact

4) Compliance:[2]

Compliance[2] (C) measures the adherence of an AI system to regulatory, ethical, and industry standards. It ensures that the system operates within the defined legal and ethical frameworks, thereby safeguarding users' rights and societal norms.

Terms in Compliance

a) Compliance Score (C)[2] :

A normalized or discrete score indicating the degree of compliance.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

b) Binary/Discrete Scoring:

C=1: Fully compliant with all regulations.

C=0.5: Partially compliant, meeting some but not all requirements.

C=0: Non-compliant.

c) Checklist-Based Assessment[2] :

A list of compliance requirements (e.g., GDPR, CCPA, ISO standards) is evaluated.

Compliance is calculated as the fraction of requirements met:

C=Number of Compliant Items/Total Number of Items

5) Explanatory Effectiveness

Explanatory Effectiveness[1] (E) evaluates how well an AI system provides explanations for its decisions. It measures two key aspects: Coverage[1]and Effectiveness[1].

Terms in Explanatory Effectiveness

a) Tool Coverage (Coverage)[1]:

The proportion of decisions for which the system provides explanations.

Formula: Coverage=Decisions with ExplanationsTotal Decisions

b) Tool Effectiveness (Ef)[1]:

Evaluates the quality of the explanations, typically using surveys, expert feedback, or clarity metrics.

Scored on a scale from 0 to 1 based on responses or evaluations.

c) Overall Explanatory Effectiveness (E)[1]:

Combines Coverage and Effectiveness into a single metric.

Formula: E=0.5·Coverage+0.5·Effectiveness

6) Accountability Score Calculation

Objective:

Aggregate the five parameters into a single accountability score.

Steps:

Formula:

- A=w_F*F+w_R*R+w_I*I+w_C*C+w_E*E
- wF,wR,wI,wC,wE Weights for each parameter (sum to 1).
- Each parameter (F,R,I,C,E) is normalized to [0, 1].

Normalization: Ensure A is within [0, 1]:

A=min(1,max(0,A))

Output:

The accountability score provides a holistic assessment ofthe AI model.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

REMAINING GAPS AND CHALLENGES

Incorporation of Impact: While societal and environmental impacts are significant, these remain poorly quantified in technical methodologies. Tools measuring long-term effects of AI decisions require further development.

Legal and Ethical Compliance: Techniques directly connecting accountability systems to compliance with diverse legal standards (GDPR, algorithmic accountability laws) need formalization and extensibility.

Scalability to Complex Models: Applying accountability methods to deep learning and multimodal models (e.g., transformers, generative AI) remains a challenging and underexplored field. Hybrid techniques like fairness-aware explanations may need adaptation for these models.

Dynamic Systems: Existing methods do not robustly address accountability in evolving AI models, such as those used in reinforcement learning or adaptive environments subject to continual learning.

CONCLUSION

The paper concludes by presenting the FRICE framework as a comprehensive model for evaluating AI accountability across five key dimensions: Fairness, Robustness, Impact, Compliance, and Explanatory Effectiveness. The framework ensures transparency, reliability, and ethical governance in AI systems by employing structured mathematical methods and weighted aggregation techniques. The proposed accountability score computation enables reproducibility and adaptability for diverse applications, addressing societal, legal, and ethical concerns while fostering trust, mitigating biases, and promoting equitable and responsible AI adoption. This framework emphasizes the integration of ethics into AI design, ensuring compliance with legal standards and clarity in decision-making processes.

REFERENCES

[1] Sharma, Shubham, Jette Henderson, and Joydeep Ghosh. "CERTIFAI: A common framework to provide explanations and analyse the fairness and robustness of black-box models." In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 166-172. 2020.

[2] De, Arkadipta, Satya Swaroop Gudipudi, Sourab Panchanan, and Maunendra Sankar Desarkar. "ComplAI: Theory of A Unified Framework for Multi-factor Assessment of Black-Box Supervised Machine Learning Models." *arXiv preprint arXiv:2212.14599* (2022).

[3] Small, Edward A., Wei Shao, Zeliang Zhang, Peihan Liu, Jeffrey Chan, Kacper Sokol and Flora Salim. “How Robust is your Fair Model? Exploring the Robustness of Prominent Fairness Strategies.” (2023).

[4] Ratz, Philipp, Franccois Hu and Arthur Charpentier. “Fairness Explainability using Optimal Transport with Applications in Image Classification.” (2023).

[5] Zhao, Yuying, Yu Wang, and Tyler Derr. "Fairness and explainability: Bridging the gap towards fair model explanations." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 9, pp. 11363-11371. 2023.

[6] Plecko, Drago, and Elias Bareinboim. "Causal fairness analysis." arXiv preprint arXiv:2207.11385 (2022).

[7] Celdrán, Alberto Huertas, Jan Kreischer, Melike Demirci, Joel Leupp, Pedro Miguel Sánchez Sánchez, Muriel Figueredo Franco, Gérôme Bovet, Gregorio Martínez Pérez and Burkhard Stiller. “A Framework Quantifying Trustworthiness of Supervised Machine and Deep Learning Models.” SafeAI@AAAI (2023).

[8] Wei, Wenqi, and Ling Liu. "Trustworthy distributed ai systems: Robustness, privacy, and governance." ACM Computing Surveys (2024).

[9] Begley, Tom, Tobias Schwedes, Christopher Frye and Ilya Feige. “Explainability for fair machine learning.” ArXiv abs/2010.07389 (2020): n. pag.

[10] Minkkinen, Matti, Joakim Laine, and Matti Mäntymäki. "Continuous auditing of artificial intelligence: A conceptualization and assessment of tools and frameworks." Digital Society 1, no. 3 (2022): 21.

[11] Galinkin, Erick. "Towards a responsible AI development lifecycle: Lessons from information security." arXiv preprint arXiv:2203.02958 (2022).

Research Article

[12] Zhang, Junzhe, and Elias Bareinboim. "Fairness in decision-making—the causal explanation formula." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1. 2018.
[13] Baniecki, Hubert, Wojciech Kretowicz, Piotr Piątyszek, Jakub Bożydar Wiśniewski and P. Biecek. “dalex: Responsible Machine Learning with Interactive Explainability and Fairness in Python.” J. Mach. Learn. Res. 22 (2020): 214:1-214:7.
[14] Khan, Masood M., and Jordan Vice. "Toward accountable and explainable artificial intelligence part one: theory and examples." IEEE Access 10 (2022): 99686-99701.
[15] Pradhan, Romila, Jiongli Zhu, Boris Glavic, and Babak Salimi. "Interpretable data-based explanations for fairness debugging." In Proceedings of the 2022 International Conference on Management of Data, pp. 247-261. 2022.
[16] Maneriker, Pranav, Codi Burley, and Srinivasan Parthasarathy. "Online fairness auditing through iterative refinement." In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 1665-1676. 2023.
[17] Makhlouf, Karima, Sami Zhioua and Catuscia Palamidessi. “Survey on Causal-based Machine Learning Fairness Notions.” ArXiv abs/2010.09553 (2020): n. pag.
[18] Mutlu, Ece Çiğdem, Niloofar Yousefi, and Ozlem Ozmen Garibay. "Contrastive counterfactual fairness in algorithmic decision-making." In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, pp. 499-507. 2022.
[19] Taka, Evdoxia, Yuri Nakao, Ryosuke Sonoda, Takuya Yokota, Lin Luo and Simone Stumpf. “Human-in-the-loop Fairness: Integrating Stakeholder Feedback to Incorporate Fairness Perspectives in Responsible AI.” (2023).
[20] Nakao, Yuri, Simone Stumpf, Subeida Ahmed, Aisha Naseer, and Lorenzo Strappelli. "Toward involving end-users in interactive human-in-the-loop AI fairness." ACM Transactions on Interactive Intelligent Systems (TiiS) 12, no. 3 (2022): 1-30.
[21] Favier, Marco, Toon Calders, Sam Pinxteren, and Jonathan Meyer. "How to be fair? a study of label and selection bias." Machine Learning 112, no. 12 (2023): 5081-5104.
[22] Guo, Hangzhi, Pranav Narayanan Venkit, Eunchae Jang, Mukund Srinath, Wenbo Zhang, Bonam Mingole, Vipul Gupta, Kush R. Varshney, S. Shyam Sundar, and Amulya Yadav. "Hey GPT, Can You be More Racist? Analysis from Crowdsourced Attempts to Elicit Biased Content from Generative AI." arXiv preprint arXiv:2410.15467 (2024).
[23] Luo, Lin, Yuri Nakao, Mathieu Chollet, Hiroya Inakoshi, and Simone Stumpf. "EARN Fairness: Explaining, Asking, Reviewing and Negotiating Artificial Intelligence Fairness Metrics Among Stakeholders." arXiv preprint arXiv:2407.11442 (2024).
[24] Leschanowsky, Anna, and Sneha Das. "Examining the interplay between privacy and fairness for speech processing: A review and perspective." arXiv preprint arXiv:2408.15391 (2024).
[25] Malik, AL-Essa, Giuseppina Andresini, Annalisa Appice, and Donato Malerba. "An XAI-based adversarial training approach for cyber-threat detection." In 2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), pp. 1-8. IEEE, 2022
[26] Stumpf, Simone, Evdoxia Taka, Yuri Nakao, Lin Luo, Ryosuke Sonoda, and Takuya 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, pp. 523-527. 2024.