Bridging the Gap between Theory and Practice: Towards Responsible AI Evaluation

Sabrina Göllner^{1,*}, Marina Tropmann-Frick^{1,*}

¹Department of Computer Science, Hamburg University of Applied Sciences, Germany

Abstract

The growing integration of artificial intelligence (AI) in diverse sectors underscores the need for comprehensive and standardized approaches to ensure AI responsibility. However, the absence of a holistic framework to evaluate the fairness, privacy-preserving, secure, explainable, and human-centered facets of AI systems poses a challenge. Addressing this gap, this research paper presents a novel approach to assessing Responsible AI by combining insights from a systematic literature review with a practical evaluation framework. The paper provides a concise overview of the key aspects of Responsible AI and highlights the findings from the literature review. Furthermore, the paper introduces a set of evaluation metrics specifically designed for the current state of the art, using different model types and data from the healthcare domain. The framework supports the evaluation of Natural Language Processing (NLP), Computer Vision (CV), and tabular data models for classification tasks. Additionally, the paper briefly demonstrates VERIFAI, an example implementation of the framework, which serves as a comprehensive tool for assessing the responsibility of AI systems. The overall objective of this research is to make a meaningful contribution to the Responsible AI discourse, providing researchers and practitioners with a valuable resource to enhance the overall responsibility of their AI systems.

Keywords

Responsible AI, Trustworthy AI, Ethical AI, Secure AI, Privacy-preserving AI, Explainable AI, Evaluation, Framework

1. Introduction

The rapid progress and extensive integration of artificial intelligence (AI) systems across diverse sectors have heightened concerns regarding their security, explainability, privacy, and fairness. Moreover, AI is becoming increasingly ingrained in daily life, leading to discussions about the roles of technologies like ChatGPT as artificial generators of text, code, and more. Ensuring Responsible AI (RAI for short) practices is crucial to maintaining trust in these systems and mitigating potential negative consequences. The relevance of RAI is not limited to technological domains. In the humanities, scholars in history, literature, and philosophy increasingly rely on AI tools for research. Beyond academia, industries like healthcare, finance, and automotive also emphasize RAI to ensure the privacy, safety, fairness, and explainability of AI-driven diagnostics, financial predictions, and autonomous driving systems, respectively.

In an effort to systematize the approach to RAI and address a noticeable gap in the literature, this

Sabrina.goellner@haw-hamburg.de (S. Göllner); marina.tropmann-frick@haw-hamburg.de (M. Tropmann-Frick) D 0000-0002-1817-7440 (S. Göllner); 0000-0003-1623-5309 (M. Tropmann-Frick)





Workshop at KI2023: 46th German Conference on Artificial Intelligence, 26 - 29 September 2023: Berlin, Germany *Corresponding author.

paper introduces a comprehensive framework for the evaluation of AI systems. This framework incorporates insights from our prior research in [1] concerning RAI characteristics and a detailed literature review [2]. The literature review attempts to provide a precise definition of RAI:

"RAI is **human-centered** and ensures users' **trust** through **ethical** ways of decision making. The decision-making must be fair, accountable, not biased, with good intentions, non-discriminating, and consistent with societal laws and norms. Responsible AI ensures, that automated decisions are **explainable** to users while always preserving users **privacy** through a **secure** implementation."

Our initial framework, VERIFAI [3], was tailored for the evaluation of Computer Vision models, utilizing specific metrics. The incorporation of these metrics is of utmost importance as they provide quantifiable measures to systematically assess the attributes of models, determining their adherence to standards of explainability, security, privacy, and fairness.

Recognizing the limitations of a singular model-centric framework and the need to have a broader evaluation spectrum, VERIFAI has been updated to include evaluations of both NLP and tabular models using appropriate metrics. This expansion ensures a more comprehensive and versatile evaluation system that can assess various AI model architectures, catering to the diversified requirements of the AI community.

The goal is to bridge the gap between theoretical understanding and practical implementation of RAI principles by providing a solid foundation for further development and refinement.

The paper is organized as follows: Section 2 summarizes the literature review findings on RAI aspects, offering a solid foundation for developing the evaluation framework. Section 3 introduces the metrics and evaluation process for specific model types and tasks. Section 4 briefly delves into the case studies to demonstrate the practical application of the framework on Computer Vision and Natural Language Processing (NLP) models as well as tabular models based on three exemplary data sets. Finally, Section 5 concludes the paper and outlines future research directions.

2. Literature Insights: RAI Aspects

The insights presented in this section are drawn from our foundational work on Responsible AI (RAI) as detailed in [1] and [2]. We provide a concise summary to understand the critical aspects that inform our current research trajectory.

Trustworthiness Trustworthiness in AI is predominantly based on user perceptions of system reliability. Key factors include prioritizing data protection, providing accurate predictions under varying conditions, and ensuring transparent explanations. Furthermore, designing systems to be user-centric and adhering to intended application behavior is essential to foster a sense of utility and fairness.

Ethics Fairness stands out as a crucial ethical requirement for AI. It is of high importance for AI systems to operate without bias and discrimination. Other significant aspects include transparency in decision justifications, alignment with Sustainable Development Goals, and strict compliance with legal standards.

Privacy Ensuring privacy, especially when handling sensitive data, is of paramount importance. Compliance with standards such as the General Data Protection Regulation (GDPR) is vital. Methods like Federated Learning, which aims to decentralize data processing, offer potential solutions, but they need to align with organizational strategies for robust data protection. Security threats in machine learning include stealing the model or sensitive user information, reconstruction attacks, and membership inference attacks, with the latter being a rapidly evolving research branch.

Security Several security threats exist within the branch of AI that warrant attention. These threats encompass poisoning attacks, where the training data is manipulated to undermine the model's performance, and adversarial attacks, in which adversaries craft malicious input samples to deceive the model and induce incorrect predictions. Addressing these concerns is an active area of research, and countermeasures are continually being developed.

Explainability The inherent opaqueness of some AI models has necessitated the development of Explainable AI (XAI). A human-centered approach to XAI emphasizes tailoring explanations to cater to diverse kinds of users. This not only ensures transparency but also empowers users in their decision-making. An intuitive user interface and visually understandable language enhance comprehension and engagement. Explainability serves as both a functional and non-functional requirement, emphasizing the AI system's inner workings and effective communication.

Human-centeredness Prioritizing human-centeredness is fundamental in RAI. The integration of human feedback and inputs, as observed in the Human-in-the-loop (HITL) approach, ensures that AI technologies align with human values while remaining efficient and user-friendly. This approach ensures that AI technologies are not solely reliant on algorithms but also benefit from human knowledge, experience, and intuition, allowing AI systems to better align with human values, expectations, and ethical considerations.

Interdependence of RAI Aspects The different aspects of RAI are closely interrelated. While ethical considerations are pivotal, they must coexist with technical requirements like security and explainability. The holistic approach encompasses both system-side and developer-side perspectives, ensuring a comprehensive RAI framework. Explainability methods must also respect privacy and security, as they are interconnected. Human-centered AI and the HITL approach play crucial roles by including human expertise and perspective. As a dynamic and interdisciplinary process, RAI requires attention and care throughout the entire system lifecycle. In the subsequent sections, we will delve deeper into the metrics of our RAI framework, building on these foundational insights.

3. RAI Evaluation Metrics

In the following, we focus on specific sub-aspects (see second row of Fig.1), which were selected, because they were identified as most important in the findings of our systematic literature review: The quantitative evaluation of Explainability, Fairness, Adversarial Robustness, and Privacy Leakage. In future work, we will expand this set of metrics.



Figure 1: RAI Evaluation Metrics

In the first iteration of our framework, our attention is centered on models for classification problems across different domain types, such as language, image, or tabular data. As illustrated in Figure 1, these are grouped into four main aspects. The sub-aspects correspond to those identified as most important in the literature analysis, and below them are the metrics already implemented in our framework.

The metrics of *Monotonicity*[4] and *Faithfulness*[5] were chosen for the Explainability assessment on the tabular model. These metrics help to evaluate the influence of individual attributes on the performance of the predictive models and understand how each attribute contributes to model performance.

As visual explanations for Neural Networks alone are often insufficient and to provide a comprehensive evaluation of explainability, four metrics were chosen, which measure explanations from different perspectives: The *Robustness* of the explanation was measured using *Max Sensitivity*[6], to ensure the explanations are stable to minor input perturbations. *Complexity* was measured using the *Sparseness*[7] metric, to see if explanations are based on a small number of features. *Faithfulness* of an explanation, which measures whether the explanations capture relevant features, was measured through the *Faithfulness Correlation*[8], and the *Randomness* of the method using *Random Logit*[9], measuring the effect of increasingly randomized parameters on the quality of explanations. Unfortunately, the metrics for assessing the explainability on NLP-models are not yet available in the current iteration.

The Fairness metrics encompass various aspects, such as *Group Fairness* (e.g. Statistical Parity Difference[10]), *Individual fairness* (e.g. Between Group Entropy Error[11]), *Data Fairness* (e.g. Prevalence of Privileged Class)[12], and *Model Performance* (e.g. F-1 Sore) providing a well-rounded evaluation.

The Robustness metrics of the framework were carefully chosen to provide a comprehensive security assessment. For tabular models, the *Zeroth Order Optimization Attack (ZOA)*[13] was selected for its versatility and effectiveness against non-differentiable models like random forests. For CV models, various adversarial attack strategies, including *FGSM*[14], *PGD* [15], *DeepFool*[16], and *Additive Uniform Noise attacks(AUN)*[14], were employed to test the model's robustness. For NLP models, *TextBugger*[17], *DeepWordBug*[18], *TextFooler*[19], and *Probability Weighted Word Saliency (PWWS)*[20] metrics were chosen to address different aspects of adversarial attacks on NLP models, providing a thorough evaluation of the model's security against various attack strategies.

For evaluating Privacy Leakage we tested them using several Membership Inference Attack (MIA) Approaches for each model. The *Black-Box-MIA*[21] metric was selected for the Tabular model, while for the neural networks (NLP and CV) we employed the metrics of *MIA via Shadow Models*[22] and *MIA via Population Data*[22]. These metrics enable the assessment of potential privacy risks in the models by quantifying if some data points of the training data can be reconstructed.

In the current development stage, we compute the *Responsibility Score* by averaging the fairness, privacy-leakage, robustness, and explainability scores, which are calculated using the averages or worst-case outcomes of their respective metrics. This method assumes an equal contribution from each category to the model's overall score. The resulting Responsibility Score is then expressed as a percentage, reflecting the model's responsibility level according to the assessed metrics.

4. Implementation and Case Studies

This section presents a part of the current implementation of the framework. Our use cases are designed to assess Responsible AI principles and specifically focus on concerns within the healthcare domain. We selected three use cases to demonstrate the framework's applicability across different tasks and data types: Detecting Skin Cancer, Recommending Medicines based on Sentiment Analysis, and Detecting Heart Diseases.

To train the models for these use cases, we used datasets from tabular data (Heart Disease Dataset[23]), image data (HAM10000 dataset[24]), and text data (Drug Review Dataset[25]). We chose representative models for each task: Random Forest[26] for tabular data, Xception[27] for image classification, and DistilBERT[28] for text classification. These models were selected based on their compatibility with evaluation metrics used in the framework, ensuring a comprehensive evaluation of the system's effectiveness.

To give an insight into the application by showing the evaluation of *adversarial robustness* of the three different models mentioned in section 4 we want to provide a few screenshots of the application in the following:



(a) Evaluation of the CV- model. The metric result (b) Evaluation of the NLP- model (without page and explanation are shown in the page header. header)



(c) Adversarial examples on tabular data with difference vectors (without page header)

Figure 2: VERIFAI Screenshots: Evaluating Adversarial Robustness of models.

The evaluation of our CV-model for detecting skin cancer is depicted in Figure 2a. The image illustrates our process of testing the model's robustness by applying various adversarial attacks, including four previously mentioned algorithms, with the aim of deceiving the model and inducing erroneous predictions. This process involves introducing perturbations to the input images and measuring the model's accuracy after each round of increasing disturbance. We find that our model has a robustness score of 0.53, indicating its vulnerability to adversarial attacks. This susceptibility means that the model can be deceived with minimal pixel changes, resulting in an 'Accuracy under Attack' of 52% using the FGSM Algorithm, thus deriving a privacy score of 5 out of 10.

Our Natural Language Processing model, as referenced in Figure 2b, is not immune to such attacks either. The adverse consequences of deceiving this model can be severe, such as the

potential for incorrect medication recommendations. We utilize four distinct algorithms, as mentioned in 3, that modify the input text to induce inaccuracies.

The table indicates the following insights: The attack was performed on 200 examples, of which the model initially misclassified 8. This resulted in an original accuracy of 92.0%. We subsequently ran the adversarial attack process on the remaining examples to derive a valid adversarial perturbation for each. 16 of these attacks failed, yielding a success rate of 82.61%. This translates to the model correctly predicting and resisting attacks on 16 out of the total 200 samples, thereby achieving an accuracy under attack of 16.0%. For the successful attacks, on average, 10.69% of words were altered to achieve a prediction change.

The model's robustness score is computed by considering the after-attack accuracy, the inverse of the adversarial attack success rate, and the degree of input modification required for a successful attack. A high robustness score corresponds to a model that resists adversarial attacks effectively, maintains high accuracy, and demands significant alterations by the adversary to be deceived. Unfortunately, in our case, the model had a robustness score of just 2 out of 10 using the most successful attacks, suggesting a need for mitigation strategies.

Next, we attempt to assess the tabular model. Figure 2c provides a visualization of the adversarial examples, enabling a deeper understanding of how the attacks influence the model's decision-making. The image illustrates features such as age versus maximum heart rate, using black lines to denote the difference vectors between the original and manipulated data points. These visual cues help to interpret the extent of data manipulation by adversarial attacks.

Despite these vulnerabilities, the model shows considerable overall robustness, scoring an encouraging 8 out of 10 (or 80%). More comprehensive results, including those on privacy leakage, fairness, and explainability, have been omitted due to paper length constraints but are available in our demo version on our project website¹.

5. Conclusion and Future Work

In conclusion, determining the level of responsibility in AI systems is a complex task that goes beyond the realm of computational metrics alone. It involves subjective considerations and requires the incorporation of human feedback and human-centered approaches. By taking into account the input and perspectives of the target audience, we can gain a more comprehensive understanding of what responsible AI entails. While our presented framework addresses various aspects of Responsible AI based on a systematic literature review, we acknowledge that it is a work in progress and has certain limitations. We recognize the need for further refinements and expansions to enhance its effectiveness. The current iteration of the framework serves as a solid foundation for evaluating Responsible AI. However, we are aware that there are specific model types, such as generative models like the GPT architecture, that may require additional attention and consideration. Additionally, as AI is applied across different domains, the framework will need to adapt and incorporate domain-specific considerations to ensure its relevance and applicability. As we continue to refine and expand our framework, we aim to address these challenges and further improve its capabilities to create a more comprehensive and robust approach to assessing the responsibility of AI systems.

¹https://www.verifai.science

References

- B. Brumen, S. Göllner, M. Tropmann-Frick, Aspects and Views on Responsible Artificial Intelligence, in: Machine Learning, Optimization, and Data Science: 8th International Workshop, LOD 2022, Certosa di Pontignano, Italy, September 19–22, 2022, Revised Selected Papers, Part I, Springer, 2023, pp. 384–398.
- [2] S. Göllner, M. Tropmann-Frick, B. Brumen, Towards a Definition of a Responsible Artificial Intelligence, in: Proceedings of the 33rd International Conference on Information Modelling and Knowledge Bases EJC 2023, University of Maribor, University Press, 2023. doi:10.18690/um.feri.5.2023.2.
- [3] S. Göllner, M. Tropmann-Frick, VERIFAI A Step Towards Evaluating the Responsibility of AI-Systems, in: B. König-Ries, S. Scherzinger, W. Lehner, G. Vossen (Eds.), BTW 2023, Gesellschaft für Informatik e.V., 2023. doi:10.18420/BTW2023-63.
- [4] R. Luss, P.-Y. Chen, A. Dhurandhar, P. Sattigeri, Y. Zhang, K. Shanmugam, C.-C. Tu, Leveraging latent features for local explanations, 2021. arXiv:1905.12698.
- [5] D. Alvarez-Melis, T. S. Jaakkola, Towards Robust Interpretability with Self-Explaining Neural Networks, 2018. URL: http://arxiv.org/abs/1806.07538, arXiv:1806.07538 [cs, stat].
- [6] C.-K. Yeh, C.-Y. Hsieh, A. S. Suggala, D. I. Inouye, P. Ravikumar, On the (In)fidelity and Sensitivity for Explanations, 2019. URL: http://arxiv.org/abs/1901.09392, arXiv:1901.09392 [cs, stat].
- [7] P. Chalasani, J. Chen, A. R. Chowdhury, S. Jha, X. Wu, Concise Explanations of Neural Networks using Adversarial Training, 2020. URL: http://arxiv.org/abs/1810.06583, arXiv:1810.06583 [cs, stat].
- [8] U. Bhatt, A. Weller, J. M. F. Moura, Evaluating and Aggregating Feature-based Model Explanations, 2020. URL: http://arxiv.org/abs/2005.00631, arXiv:2005.00631 [cs, stat].
- [9] L. Sixt, M. Granz, T. Landgraf, When Explanations Lie: Why Many Modified BP Attributions Fail, 2020. URL: http://arxiv.org/abs/1912.09818, arXiv:1912.09818 [cs, stat].
- [10] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A Survey on Bias and Fairness in Machine Learning, 2021. URL: https://doi.org/10.1145/3457607. doi:10.1145/3457607.
- [11] T. Speicher, H. Heidari, N. Grgic-Hlaca, K. P. Gummadi, A. Singla, A. Weller, M. B. Zafar, A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual &Group Unfairness via Inequality Indices, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, London United Kingdom, 2018, pp. 2239–2248. URL: https://dl.acm.org/doi/10.1145/3219819.3220046. doi:10.1145/ 3219819.3220046.
- [12] C. Allen, M. Ahmad, C. Eckert, J. Hu, V. Kumar, A. Teredesai, fairmlhealth: Tools and tutorials for fairness evaluation in healthcare machine learning., https://github.com/ KenSciResearch/fairMLHealth, 2020.
- [13] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, C.-J. Hsieh, Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models, in: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 15–26. URL: https://doi.org/10.1145/3128572.3140448. doi:10.1145/3128572.3140448.
- [14] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples,

arXiv preprint arXiv:1412.6572 (2014).

- [15] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, arXiv preprint arXiv:1706.06083 (2017).
- [16] S.-M. Moosavi-Dezfooli, A. Fawzi, P. Frossard, Deepfool: a simple and accurate method to fool deep neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2574–2582.
- [17] J. Li, S. Ji, T. Du, B. Li, T. Wang, Textbugger: Generating adversarial text against real-world applications, arXiv preprint arXiv:1812.05271 (2018).
- [18] J. Gao, J. Lanchantin, M. L. Soffa, Y. Qi, Black-box generation of adversarial text sequences to evade deep learning classifiers, in: 2018 IEEE Security and Privacy Workshops (SPW), IEEE, 2018, pp. 50–56.
- [19] D. Jin, Z. Jin, J. T. Zhou, P. Szolovits, Is bert really robust? a strong baseline for natural language attack on text classification and entailment, in: Proceedings of the AAAI conference on artificial intelligence, volume 34, 2020, pp. 8018–8025.
- [20] S. Ren, Y. Deng, K. He, W. Che, Generating natural language adversarial examples through probability weighted word saliency, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1085–1097. URL: https://aclanthology.org/P19-1103. doi:10.18653/v1/P19-1103.
- [21] R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership Inference Attacks Against Machine Learning Models, in: 2017 IEEE Symposium on Security and Privacy (SP), IEEE, San Jose, CA, USA, 2017, pp. 3–18. URL: http://ieeexplore.ieee.org/document/7958568/. doi:10.1109/SP.2017.41.
- [22] J. Ye, A. Maddi, S. K. Murakonda, V. Bindschaedler, R. Shokri, Enhanced Membership Inference Attacks against Machine Learning Models, 2022. URL: http://arxiv.org/abs/2111. 09679, arXiv:2111.09679 [cs, stat].
- [23] A. Janosi, W. Steinbrunn, M. Pfisterer, R. Detrano, Uci machine learning repository-heart disease data set, School Inf. Comput. Sci., Univ. California, Irvine, CA, USA (1988).
- [24] P. Tschandl, C. Rosendahl, H. Kittler, The ham10000 dataset, a large collection of multisource dermatoscopic images of common pigmented skin lesions, Scientific data 5 (2018) 1–9.
- [25] F. Gräßer, S. Kallumadi, H. Malberg, S. Zaunseder, Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning, in: Proceedings of the 2018 International Conference on Digital Health, DH '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 121–125. URL: https://doi.org/10.1145/3194658. 3194677. doi:10.1145/3194658.3194677.
- [26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, Journal of machine learning research 12 (2011) 2825–2830.
- [27] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
- [28] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).