MARTA

Machine-learning Auditing for Robust and Transparent (Public) Administration

Mariana Pinto[1] Matilde Silva[1] Matilde Ferreira[1] Beatriz Félix[1]
Iakovina Kindylidi[2,3] Marília Barandas[1] Hugo Gamboa[1,4] André Carreiro[1]

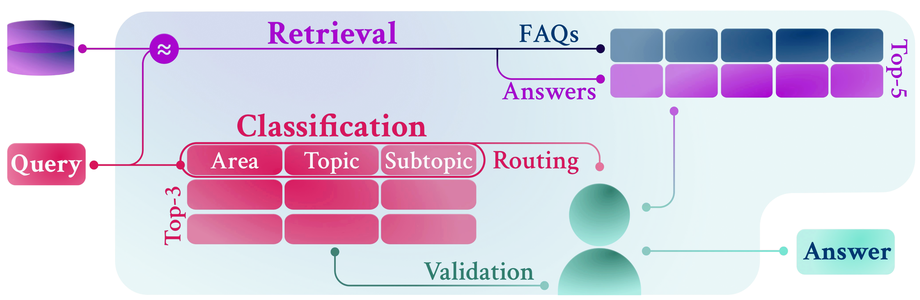
[1] Fraunhofer Portugal AICOS, Porto, Portugal
[2] NOVA School of Law, Lisbon, Portugal
[3] Vieira de Almeida Associados (VdA), Lisbon, Portugal
[4] Laboratory of Instrumentation, Biomedical Engineering, Radiation Physics (LIBPhys-UNL), NOVA SST, Caparica, Portugal.

Current auditing frameworks provide organizational guidance but lack empirical tools for operational audits in public-sector AI systems [1,2].

MARTA is a **multidisciplinary framework** that audits AI systems by **linking empirical model testing with legal and human-oversight requirements.**

Through the **operationalization of HUDERIA framework**, applied to the Smart e-Desk system, MARTA provides **concrete guidelines for providers and deployers**

1. Context-Based Risk Analysis (COBRA)



The service uses AI models to **assist human operators in responding to taxpayer queries**, aiming to improve both the efficiency and quality of interactions.

Additionally to the **Topic Routing** and **Suggestion Retrieval** modules, the framework incorporates a **Keyphrase Identification** and an **Anonymization Module**.

Based on EU AI Act, the risk classification is **minimal-risk**.

2. Stakeholder Engagement Process (SEP)

AXIANS → **Provider** : Develops and supplies the AI system

Portuguese Tax Authority → **Deployer** : Integrating the AI system into its public-facing services and operator workflows

Stakeholders contact is kept through **active communication** supported by **structured questionnaires and interviews**.

3. Risk and Impact Assessment (RIA)

	Risk	Audit
Bias	Unequal Treatment based on gender	Counterfactual Analysis (gender substitutions)
Robustness	Subpar performance for queries with realistic errors	Text Perturbation Attacks
Privacy	Data Leakage	Text Attacks against the Anonymizer
XAI	Misleading/Unhelpful Explanations	Text Attacks against the Keyphrases Extractor
Human Agency	Under or Over Reliance in the System	Confidence Analysis and review of pilot protocol

Table 1: Summary to the risk analysis conducted to the Smart e-Desk system and chosen technical audit based on this risk mitigation goal.

4. Mitigation Plan (MP)

	Providers	Deployers
Bias	**(S)** Counterfactual augmentation as a mitigation measure	**(G)** Monitor demographic proxies for fairness drift; **(S)** Periodic linguistic audits to identify representational and allocation harms.
Robustness	**(G)** Provide robustness documentation; **(S)** Perturbation based data augmentation for training.	**(G)** Adversarial testing integrated into lifecycle risk management.
Privacy	**(G)** Strict data minimization in the AI and data lifecycle; **(S)** Finetune anonimizer with adversarial attacks.	**(G)** Regular leakage audits and monitoring.
XAI	**(G)** Include model-internal explanation methods; **(G)** Contextualization warnings in documentation.	**(G)** Keep updated online policies (TC, privacy).
Human Agency	**(S)** Confidence-based AI-aid reject option.	**(G)** Operator training and intervention traceability; **(G)** Periodic review of abstention thresholds; **(S)** Monitor the operators satisfaction with the system.

Table 2: Recommendations for providers and deployers, classified based on specific (S) to the use case or generic (G) to any system.

Conclusions

MARTA provides empirical evidence directly aligned with HUDERIA's normative commitments to non-discrimination, transparency, privacy, and accountable administration, through a concise legal, technical and socio-technical evaluation. **Based on the Smart e-Desk use case, MARTA can faithfully uncover both strengths and limitation**:

+ Negligible bias
+ Low susceptibility to linguistic perturbations
+ Well-calibrated uncertainty
- Unreliable anonymization
- Partial explanation sufficiency
- Noted oversight challenges and tensions

Future work includes auditing the system's retrieval model, and expand and refine the employed methods for agnostic integration in other use cases.

References

[1] Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P. 2020. In Proceedings of the 2020 conference on fairness, accountability, and transparency, 33–44.

[2] NIST. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Technical report, National Institute of Standards and Technology.

Acknowledgements

The authors acknowledge the Portuguese Tax Authority, Autoridade Tributária e Aduaneira (AT), and AXIANS for their essential collaboration.

This work was supported by Fundação para a Ciência e Tecnologia (FCT) under the project "Machine Learning Auditing for Robust and Transparent Administration (MARTA)" [2024.07646.IACDC/2024], funded through the *Inteligência Artificial, Ciência dos Dados e Cibersegurança de relevância na Administração Pública* call.

Read the Paper!