Specifying Fairness and Transparency Requirements for Public Benefit Allocation

Amanda Aline F.C. Vicenzi[1][0009−0005−3971−9148], José Siqueira de Cerqueira[2][0000−0002−8143−1042], Pekka Abrahamsson[2][0000−0002−4360−2226], and Edna Dias Canedo[1][0000−0002−2159−339X]

[1] University of Brasília (UnB), Department of Computer Science, Brasília–DF, Brazil
amandaaline3@gmail.com, ednacanedo@unb.br
[2] Tampere University, Faculty of Information Technology and Communication Sciences, Tampere, Finland
jose.siqueiradecerqueira@tuni.fi, pekka.abrahamsson@tuni.fi

Abstract. Context and motivation: The increasing adoption of AI in public administration requires translating ethical and legal expectations into verifiable requirements. In Brazil, AI is increasingly explored to support eligibility assessment and allocation in social programs, yet there is limited guidance on how to specify and validate fairness and transparency in such high-stakes settings. **Question/problem:** This research preview investigates how fairness and transparency can be operationalized as measurable non-functional requirements (NFRs) for AI-supported social benefit allocation. **Principal ideas/results:** As a proof-of-concept, we simulate scholarship allocation using ProUni data and assess the model through predictive performance, group fairness metrics, and SHAP-based explainability evidence. Results indicate satisfactory accuracy while revealing measurable racial disparities, motivating bias mitigation. **Contribution:** We propose an initial RE-oriented framework that integrates fairness indicators, explainability artifacts, and accountability mechanisms into specification and validation checkpoints, supporting early-stage discussion and feedback on responsible AI in digital government.

Keywords: Responsible AI · Non-Functional Requirements · Fairness · Transparency · Digital Government

1 Introduction

The digitalization of public services has accelerated the adoption of Artificial Intelligence (AI) to improve efficiency, scalability, and resource allocation in public administration [12,17]. In Brazil, AI-based approaches are increasingly considered to support decision-making in social programs, including eligibility assessment and benefit allocation. However, in high-stakes public contexts, algorithmic decisions may affect citizens' rights and reproduce or amplify historical inequalities embedded in data and institutional practices [1,9,4,15]. From a software engineering perspective, these risks translate into ethical non-functional

requirements (NFRs) such as fairness, transparency, and accountability, which are often weakly specified, validated, and monitored. In practice, AI systems are frequently assessed primarily through predictive performance, while ethical properties are addressed later or informally [13], a particularly critical issue in public-sector settings subject to constitutional principles and data protection obligations.

Operationalizing ethical requirements requires measurable criteria and auditable evidence. Explainability plays a central role, as stakeholders must be able to understand and scrutinize the drivers of automated recommendations. Techniques such as SHapley Additive exPlanations (SHAP) provide global and local insights into feature influence, supporting transparency and bias diagnosis [18]. From a requirements engineering (RE) perspective, such explainability artifacts can serve as validation evidence, strengthening traceability between ethical principles, requirements, and system behaviour.

This research preview investigates how fairness and transparency can be specified as verifiable NFRs for AI-based social benefit allocation, using the University for All Program (ProUni) as an illustrative context. ProUni is a Brazilian federal scholarship initiative launched in 2004 to expand access to higher education based on socioeconomic and educational criteria. We conduct a proof-of-concept audit using a simulated predictive model inspired by ProUni data, assessing fairness through quantitative metrics and transparency through SHAP-based analyses. Based on these insights, we propose an initial RE-oriented framework that integrates ethical principles, legal constraints, and measurable fit criteria into specification and validation checkpoints, aiming to stimulate discussion on systematic approaches to ethical requirements in digital government.

2 Background and Related Work

The adoption of AI in public-sector decision-making has expanded rapidly in recent years, particularly in domains such as education, health, and social assistance [1]. While data-driven systems promise efficiency and scalability, their use in high-stakes public contexts raises persistent concerns regarding fairness, transparency, accountability, and compliance with fundamental rights [3]. When algorithmic outputs influence access to social benefits, these concerns must be addressed not only as ethical issues but also as explicit system-level requirements.

A substantial body of research has documented algorithmic bias as a central risk in AI-supported decision systems, especially when historical data encode structural inequalities [8]. In public-sector applications, biased outcomes may directly undermine principles of equality, impersonality, and legality. To address these risks, prior work on fairness aware AI proposes quantitative metrics such as Statistical Parity Difference, Equal Opportunity Difference, and Disparate Impact to assess group-level disparities [10]. Toolkits such as IBM AIF360 [2] operationalize these metrics; however, their use is typically detached from RE artifacts, such as fit criteria, validation checkpoints, or traceability mechanisms. Complementary to fairness assessment, explainable AI (XAI) techniques have

been proposed to enhance transparency and trust in algorithmic decisions. Methods such as SHapley Additive exPlanations (SHAP) provide global and local interpretability, enabling stakeholders to inspect feature influence and diagnose potential bias [18]. Although XAI is often treated as a post hoc analysis mechanism, recent studies highlight its relevance for accountability and regulatory compliance. Nevertheless, the systematic use of explainability artifacts as validation evidence for ethical NFRs remains under-explored in RE research.

At the governance level, international initiatives such as the OECD AI Principles [14], the UNESCO Recommendation on AI Ethics [11], and the European Union AI Act [19] emphasize transparency, non-discrimination, accountability, and human oversight, particularly for high-risk AI systems. Despite their normative importance, these frameworks provide limited guidance on how ethical principles can be translated into verifiable system requirements or embedded into software development practices. In Brazil, this gap is reflected in the coexistence of constitutional principles, LGPD obligations, and national AI strategies that still lack concrete engineering level operationalization [16].

From a methodological perspective, many AI projects follow established data mining and machine learning lifecycles, such as CRISP-DM [5]. While such lifecycles offer structured guidance for technical activities, ethical concerns such as fairness and transparency are typically addressed implicitly within phases like business understanding or evaluation, rather than specified through measurable fit criteria or systematically validated across the lifecycle. RE research has therefore emphasized the need to operationalize ethical principles as NFRs, enabling their specification, validation, and monitoring alongside traditional quality attributes [15]. However, existing approaches often remain conceptual or focus on isolated techniques, without demonstrating integrated workflows that combine fairness metrics, explainability artifacts, and accountability mechanisms.

Our research addresses a gap at the intersection of AI lifecycles and RE. Rather than proposing a new end-to-end development process, we extend established lifecycles such as CRISP-DM with an explicit RE layer that treats fairness and transparency as verifiable NFRs. This layer is operationalized through measurable fit criteria, explainability-based validation evidence, and traceability links between ethical principles, legal norms, and system-level requirements. Grounded in a preliminary case study from Brazilian digital government, this work contributes an early-stage, RE-oriented perspective on translating ethical AI principles into actionable and auditable engineering practices.

3 Study Setting

This research preview examines whether AI-based decision-support models in public administration may reproduce or amplify social inequalities, and how ethical requirements can be operationalized and assessed within the AI lifecycle. We focus on fairness and transparency as verifiable NFRs in a digital government context, guided by the following research questions:

RQ1. To what extent can AI-based predictive models for public benefit allocation exhibit measurable disparities across social groups?

RQ2. How can fairness and transparency be specified and assessed as NFRs within a context-sensitive framework for AI development in the Brazilian public sector?

To address **RQ1**, we conducted a preliminary audit of a predictive model simulating scholarship allocation in the University for All Program (ProUni), evaluating predictive performance, explainability, and fairness indicators. To address **RQ2**, we derived an initial framework that embeds fairness and transparency as measurable NFRs within the RE process. The study follows the CRISP-DM methodology [5], extended with two RE-oriented checkpoints: **RE-1 (Specification)**, which defines explicit fairness and transparency criteria, and **RE-2 (Validation)**, which verifies compliance using quantitative fairness metrics and explainability evidence.

The dataset was obtained from the Brazilian Federal Government Open Data Portal (MEC) and contains records of ProUni scholarships awarded between 2005 and 2020. Available attributes include region, institution, field of study, modality, study shift, scholarship type, and sociodemographic variables (sex, race/colour, age, disability). As the public dataset includes only successful applicants, a synthetic set of non-beneficiaries was generated to enable supervised learning and exploratory fairness auditing. This synthetic data was used solely for proof-of-concept purposes and does not aim to reproduce real allocation outcomes. Categorical variables were one-hot encoded, and a target variable (`PROBABILIDADE_BOLSA`) was introduced to simulate approval likelihood. Data were split into training (80%) and testing (20%) sets, with numerical features standardized using `StandardScaler`. We employed `XGBRegressor` [6], selected for its robustness with tabular data.

Model performance was evaluated using MAE, MSE, RMSE, R^2, and MAPE [20]. The model achieved satisfactory predictive performance (MAE = 0.125). Explainability was assessed using SHAP, providing global and local insights into feature influence and supporting transparency oriented validation [18]. The most influential attributes included region, race/colour, scholarship type, and year of award. Group fairness was evaluated using Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD), and Average Odds Difference (AOD) [10]. The unmitigated model exhibited measurable disparities across racial groups (odds ratio = 1.04, $p < 0.001$). Applying the Reweighing technique from AIF360 reduced disparities across all metrics (SPD: –0.09 to –0.02; EOD: –0.06 to –0.01; AOD: –0.07 to –0.01). The results indicate that predictive accuracy alone is insufficient to assess AI systems in sensitive public-sector contexts. Fairness auditing revealed group-level disparities prior to mitigation, while explainability analyses supported transparency and traceable validation of ethical NFRs. These findings provide initial evidence addressing **RQ1** and inform the proposed RE-oriented framework addressing **RQ2**.

4 An Ethical Framework for AI in Government

Grounded in Article 37 of Brazil's 1988 Constitution legality, impersonality, morality, publicity, and efficiency, this framework treats fairness and transparency as first-class NFRs for AI systems in government. Rather than proposing a new AI lifecycle, it introduces an explicit RE layer that complements existing lifecycles by enabling the specification, validation, and monitoring of ethical requirements. Insights from the ProUni case illustrate that models evaluated solely through predictive performance may produce indirect discrimination, reinforcing the need to embed equity criteria from the outset and to audit them continuously. Figure 1 summarizes the framework around two RE checkpoints: **RE-1 (Specification)** and **RE-2 (Validation)**.

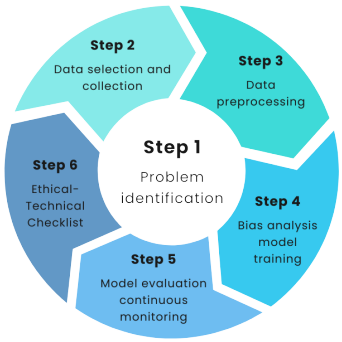

Fig. 1. Ethical framework for AI in Brazilian digital government with RE checkpoints for fairness and transparency.

Problem Identification and Data Selection. AI initiatives in public administration should begin with a contextualized problem definition, identifying affected social groups and intended policy outcomes. Ethical data selection must be guided by relevance, representativeness, and non-discrimination [7]. Sensitive attributes should be retained not removed to enable bias detection and mitigation. In the ProUni case, the availability of data only for awarded scholarships limited impartiality assessment, motivating the use of complementary public sources or controlled synthetic data strictly for exploratory fairness auditing.

Data Preprocessing. Preprocessing decisions directly affect fairness and must therefore be auditable. Public-sector datasets often contain missing or "Not Informed" values for sensitive attributes. Ethical preprocessing requires diagnosing missingness patterns, applying statistically sound treatments, and documenting all transformations. Removing sensitive attributes to "avoid bias" is counterproductive, as it prevents bias detection. Transparent documentation of cleaning and encoding decisions supports traceability and accountability.

Model Evaluation and Monitoring. Evaluation extends beyond technical performance to include verification of ethical NFRs. In the ProUni proof-of-concept, fairness auditing revealed measurable disparities across racial groups, motivating mitigation actions. Explainability techniques such as SHAP provide local and global insights into influential features, supporting transparency and managerial oversight [18]. Human oversight remains essential, as quantitative indicators must be interpreted in light of public policies and social context. Continuous monitoring and documented mitigation cycles are required to sustain legality and public trust over time.

RE Checkpoints and Fit Criteria. At **RE-1 (Specification)**, ethical requirements are defined with explicit fit criteria, such as: (i) **Fairness**: $|\text{SPD}|$, $|\text{EOD}|$, $|\text{AOD}| \leq 0.02$ for protected attributes; (ii) **Disparate Impact**: DI within $[0.8, 1.25]$; (iii) **Transparency**: provision of local SHAP-based explanations within predefined latency and fidelity thresholds; and (iv) **Auditability**: complete decision trails stored for all inferences over a defined retention period. At **RE-2 (Validation)**, compliance is verified through fairness reports, explainability artifacts, and records of mitigation and re-evaluation.

Reflective Dimensions. To avoid a purely procedural checklist, the framework incorporates five reflective dimensions guiding documentation and justification across the lifecycle: (i) algorithmic sensitivity and informational justice; (ii) epistemic transparency and cognitive accessibility; (iii) technical responsibility and accountability; (iv) social sustainability and institutional reflexivity; and (v) regulatory compliance and data governance. These dimensions promote evidence based reflection rather than binary compliance judgments. The framework broadens fairness assessment beyond statistical metrics by integrating legal, ethical, and institutional considerations into RE artifacts. In the Brazilian context where a consolidated AI law is still under discussion (Bill No. 2338/2023), it provides a structured and auditable path for aligning AI-supported public decisions with constitutional principles and fundamental rights.

5 Discussion

Performance, Fairness, and Validation. The results confirm that predictive performance alone is insufficient for evaluating AI systems in sensitive public-sector contexts. Although the proof-of-concept model achieved satisfactory accuracy, fairness auditing revealed measurable racial disparities, indicating the risk of reproducing historical inequalities through automated decisions. The reduction of these disparities after bias mitigation reinforces the need to treat fairness verification as a formal validation activity (RE-2 checkpoint), rather than as a post hoc concern. These findings highlight that accuracy and fairness must be jointly assessed throughout the AI lifecycle.

Legal, Institutional, and Managerial Implications. The observed disparities are directly related to constitutional principles of equality and non-discrimination, as well as to obligations established by the LGPD. AI-based

decision-support systems that produce group-level disparities without systematic monitoring or mitigation may conflict with the public sector's duty to prevent indirect discrimination. By enabling traceable auditing and validation of ethical requirements, the proposed framework helps translate legal and constitutional principles into operational system-level criteria, supporting both regulatory compliance and managerial oversight.

Implications for Requirements Engineering Practice. From a RE perspective, the framework operationalizes fairness, explainability, and auditability as NFRs that must be specified, validated, and monitored alongside traditional quality attributes. Fairness metrics such as SPD and EOD serve as quantitative proxies for ethical compliance, while explainability artifacts support transparency, interpretability, and debugging. Auditability, in turn, enhances accountability and maintainability by enabling traceability of model decisions and periodic review. Integrating these attributes into AI development and validation workflows strengthens trustworthiness and societal legitimacy, particularly in digital government contexts.

Positioning with Respect to Existing Frameworks. Unlike existing AI governance initiatives that primarily articulate high-level ethical principles, the proposed framework focuses on their operationalization within software engineering practice. Rather than replacing established AI lifecycles, it complements them with an explicit RE layer that provides concrete fit criteria, validation evidence, and traceability mechanisms. This positioning bridges the gap between normative ethical guidance and day-to-day engineering activities in public-sector AI development.

RQ1. Preliminary Insight. AI-based models for social benefit allocation may exhibit measurable group-level disparities when trained on historically constrained or unbalanced data, underscoring the need for explicit fairness requirements and validation checkpoints.

RQ2. Preliminary Insight. Fairness and transparency can be specified and assessed as measurable NFRs through an RE-oriented approach that combines fairness metrics, explainability artifacts, and traceability mechanisms.

6 Conclusion

This research preview examined how fairness and transparency can be operationalized as verifiable NFRs for AI-based decision-support systems in digital government. Using a proof-of-concept predictive model inspired by the CRISP-DM methodology, the study illustrated that predictive performance alone is insufficient in high-stakes public-sector contexts. The ProUni case highlighted measurable fairness disparities and showed how explainability and bias mitigation techniques can support ethical auditing of AI models.

Beyond the case illustration, the paper proposed an initial requirements oriented framework that integrates fairness and transparency into the AI lifecycle through explicit RE checkpoints for specification and validation. Grounded in Brazilian constitutional principles and data protection regulations, the framework translates abstract ethical and legal obligations into measurable criteria, validation evidence, and audit trails, complementing existing AI lifecycles rather than replacing them. As an early-stage investigation, this work does not claim definitive validation of the proposed framework. Instead, it aims to stimulate discussion within the RE community on how ethical requirements for AI systems can be systematically specified, validated, and monitored in public-sector applications. Future work will empirically evaluate the framework in real governmental settings and refine its criteria and validation mechanisms across different policy domains.

Data Availability Statement The data that support the findings of this study are openly available in Zenodo at https://zenodo.org/records/17410424.

Acknowledgments This work was supported by CONVERGENCE of Humans and Machines (220025) and the EVIL-AI "The identification and the mitigation of the negative effects of Artificial Intelligence Agents" (JAES/2024/EVIL-AI) projects by Jane and Aatos Erkko Foundation and the "Multifaceted ripple effects and limitations of human-AI interplay at work, business and society (SYNTHETICA)" project (358714) by Research Council of Finland. We thank the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Grant Nº 300883/2025-0.

References

1. de Almeida, P.G.R., dos Santos Júnior, C.D.: Artificial intelligence governance: Understanding how public organizations implement it. Gov. Inf. Q. **42**(1), 102003 (2025). `https://doi.org/10.1016/J.GIQ.2024.102003`
2. Blow, C.H., Qian, L., Gibson, C., Obiomon, P., Dong, X.: Comprehensive validation on reweighting samples for bias mitigation via aif360 (2023), `https://arxiv.org/abs/2312.12560`
3. Casillas, J.: Bias and discrimination in machine decision-making systems. Ethics of Artificial Intelligence **41**, 13–38 (2024)
4. de Cerqueira, J.A.S., Azevedo, A.P.D., Leão, H.A.T., Canedo, E.D.: Guide for artificial intelligence ethical requirements elicitation - RE4AI ethical guide. In: 55th Hawaii International Conference on System Sciences, HICSS. pp. 1–10. ScholarSpace, http://hdl.handle.net/10125/80015 (2022)
5. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R.: The crisp-dm user guide. In: 4th CRISP-DM SIG Workshop in Brussels in March. vol. 1999, pp. 1–14. sn, NCR Systems Engineering Copenhagen, https://s2.smu.edu/ mhd/8331f03/crisp.pdf (1999)
6. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. vol. 1, pp. 785–794. arXiv, https://arxiv.org/abs/1603.02754 (2016)

7. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Iii, H.D., Crawford, K.: Datasheets for datasets. Communications of the ACM **64**(12), 86–92 (2021)
8. Gentelet, K., Mizrahi, S.K.: A human-centered approach to ai governance: Operationalizing human rights through citizen participation. In: Human-Centered AI, pp. 215–230. Chapman and Hall/CRC, https://www.taylorfrancis.com/chapters/oa-edit/10.1201/9781003320791-24/human-centered-approach-ai-governance-karine-gentelet-sarit-mizrahi (2024)
9. Gonçalves, C.D., de Paoli Menescal, E., de Mendonça, F.L.L., Canedo, E.D.: Trust in AI: perspectives of c-level executives in brazilian organizations. In: Proceedings of the XXIII Brazilian Symposium on Software Quality, SBQS 2024, Salvador, Bahia, Brazil, November 5-8, 2024. pp. 147–157. ACM, https://doi.org/10.1145/3701625.3701654 (2024)
10. González-Sendino, R., Serrano, E., Bajo, J., Novais, P.: A review of bias and fairness in artificial intelligence. International Journal of Interactive Multimedia and Artificial Intelligence **9**, 1–13 (2023)
11. Kettemann, D.M.C.: Unesco recommendation on the ethics of artificial intelligence. Conditions for the Implementation in Germany **1**, 1–43 (2022)
12. Lin, K., Shen, C., Cheng, S.: Applications of AI in digital governance services for local taxes- a case of the local tax bureau of taichung city government. In: Proceedings of the 25th Annual International Conference on Digital Government Research, DGO 2024, Taipei, Taiwan, June 11-14, 2024. pp. 6–18. ACM, https://doi.org/10.1145/3657054.3657056 (2024)
13. Mellouli, S., Janssen, M., Ojo, A.: Introduction to the issue on artificial intelligence in the public sector: Risks and benefits of AI for governments. Digit. Gov. Res. Pract. **5**(1), 1:1–1:6 (2024). https://doi.org/10.1145/3636550
14. OECD: Recommendation of the council on artificial intelligence. OECD: Paris, France pp. 1–12 (2024), https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449, originally adopted on 22 May 2019; amended on 3 May 2024
15. de Paula Porto, D., Prado, R.D.C.V., dos Santos Marques, G., Serrano, A.L.M., Mendonça, F.L.L., Canedo, E.D.: Ethical requirements in the age of artificial intelligence: A systematic literature review. In: Proceedings of the 21st Brazilian Symposium on Information Systems, SBSI 2025, Recife, Brazil, May 19-23, 2025. pp. 663–672. SBC, https://doi.org/10.5753/sbsi.2025.246613 (2025)
16. Presidência da República do Brasil: Lei nº 13.709, de 14 de agosto de 2018. https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm (2018)
17. Reis, A.R., Lopes, J.M., da Costa, J.M., de Jesus, T.F., Torres, T.P.d.R.: Artificial intelligence as a tool applicable to public administration: A look at the human resources area. ARACÊ **6**(4), 18213–18238 (Dec 2024). https://doi.org/10.56238/arev6n4-422
18. Salih, A.M., Raisi-Estabragh, Z., Galazzo, I.B., Radeva, P., Petersen, S.E., Lekadir, K., Menegaz, G.: A perspective on explainable artificial intelligence methods: Shap and lime. Advanced Intelligent Systems **7**(1), 2400304 (2025)
19. Sonsini, W., Parliament, T.E.: The eu artificial intelligence act. European Union pp. 1–10 (2024), https://www.wsgr.com/a/web/qrkz1SnNzWw6nk7B3oAyDa/10-things-you-should-know-about-the-eu-artificial-intelligence-act_v2.pdf
20. Tatachar, A.V.: Comparative assessment of regression models based on model evaluation metrics. International Research Journal of Engineering and Technology (IRJET) **8**(09), 2395–0056 (2021)