Journal of Asian Studies: Culture, Language, Art and Communications
e-ISSN 2774-5554 | HAQI Publishing Service

Volume 7 Issue 1, 2026
https://journal.haqipub.com/index.php/jas/index

E-mail:
lena.schmidt@sodalab.lmu.de

Ludwig Maximilian University of Munich (LMU)
Social Data Science Lab
Germany / Europe

Assessing Algorithmic Fairness and Bias in Predictive Social Data Science Models in German Public Administration

***Lena Schmidt**

Abstract

This study investigates the potential for algorithmic bias in Machine Learning (ML) systems deployed within German public administration, specifically in social resource allocation and urban planning decision support, employing a Computational Fairness Audit methodology that integrates statistical bias measurement with sociologically grounded fairness theory. The research operates within the layered regulatory environment of the European Union's Artificial Intelligence Act, the General Data Protection Regulation, and the German Sozialgesetzbuch, constructing an audit protocol that is simultaneously technically rigorous, legally compliant, and sociologically informed. Using simulated administrative datasets and pseudo-anonymised historical records compiled in compliance with German federal data protection law, the study operationalises six socially grounded fairness metrics applied to three ML model architectures: logistic regression, gradient boosting, and a deep neural network, trained on tasks representative of Jobcenter benefit allocation scoring and municipal social housing prioritisation. Statistical measurement of implicit discrimination against minority and socially vulnerable population groups defined with reference to the protected characteristics enumerated in the Allgemeines Gleichbehandlungsgesetz reveals consistent and statistically significant fairness metric violations across all three model architectures, with gradient boosting and deep neural network models demonstrating substantially higher demographic parity gaps against Turkish-German, refugee-background, and single-parent household population subgroups than against the majority population reference group. The audit further reveals that model accuracy, as measured by standard classification performance metrics, is inversely correlated with fairness metric performance across protected subgroups. This finding directly challenges the technically unwarranted assumption prevalent in German public-sector AI procurement that high model accuracy implies equitable decision outcomes. The study contributes the Socially Grounded Algorithmic Audit Framework (SGAAF) as a replicable, legally anchored protocol for the pre-deployment and in-deployment assessment of ML systems in German public administration contexts.

Keywords: Algorithmic Fairness, Bias Audit, Machine Learning, German Public Administration, EU AI Act, GDPR, Social Resource Allocation, Computational Fairness, Demographic Parity, Discrimination.

1. INTRODUCTION

The progressive integration of Machine Learning systems into the administrative decision-making processes of German public institutions represents one of the most consequential and underscrutinised transformations in the governance of social rights in contemporary Germany. Across a range of administrative functions from the scoring of benefit entitlement applications at Jobcentern under the Sozialgesetzbuch II (SGB II) framework, to the prioritisation of applicants for social housing (Sozialwohnungen) under municipal allocation systems, to the risk stratification of child welfare cases by Jugendämter (Youth Welfare Offices) ML-based decision support tools have been incrementally adopted by federal, state, and municipal administrations attracted by the promise of administrative efficiency gains, consistency of decision-making, and the reduction of case processing backlogs that constitute a persistent structural challenge for German social service delivery (Kuhlmann & Heuberger, 2021). These adoptions have occurred within a governance environment that, until the recent entry into force of the EU Artificial Intelligence Act (EU AI Act, Regulation (EU) 2024/1689), lacked a systematic regulatory framework for the pre-deployment assessment of AI systems deployed in high-stakes public administration contexts.

The European Union's AI Act, which entered into force on 1 August 2024 and will become fully applicable in a phased schedule extending to 2027, designates AI systems used in the administration of public benefits and social services as high-risk systems under Annex III, imposing mandatory conformity assessment

requirements, human oversight obligations, transparency and logging requirements, and fundamental rights impact assessment obligations that represent the most stringent AI governance requirements applicable to any commercial or governmental AI deployment in the global regulatory landscape (European Parliament, 2024). The AI Act's high-risk designation for public administration AI is a direct legislative recognition of the concern that has animated the algorithmic fairness scholarship for over a decade: that ML systems trained on historically generated administrative datasets encode the structural inequalities, discriminatory practices, and systemic biases embedded in those datasets, and reproduce or amplify those inequalities in their outputs in ways that are not visible to human reviewers examining the model's aggregate performance metrics (Barocas & Hardt, 2017; Chouldechova, 2017; Dwork et al., 2012).

In the German public administration context, the relevant protected characteristics against which algorithmic bias must be assessed are defined by the Allgemeines Gleichbehandlungsgesetz (AGG, General Equal Treatment Act, 2006), which prohibits discrimination based on race, ethnic origin, gender, religion or belief, disability, age, and sexual orientation in the provision of goods and services, including public services. The demographic composition of German social benefit recipient populations creates a specific and politically sensitive bias risk landscape: Turkish-German citizens and residents, who constitute Germany's largest ethnic minority community with approximately 2.8 million individuals, and persons with refugee or asylum-seeker backgrounds, who have constituted a substantial share of Jobcenter caseloads since 2015-2016, are significantly over-represented among SGB II benefit recipients relative to their share of the total population (Bundesagentur für Arbeit, 2023). If an ML system trained on historical SGB II benefit allocation data learns statistical associations between ethnic or national origin proxies such as postal code, surname encoding patterns, or language of initial application and benefit award rates, it will reproduce those associations in its scoring outputs, generating discriminatory allocation recommendations that violate both the AGG and the fundamental rights protections of the EU Charter of Fundamental Rights without any discriminatory intent on the part of the model's designers or deployers.

Despite the clarity of this risk and the imminence of the EU AI Act's high-risk system requirements, published empirical research on algorithmic fairness specifically in German public administration ML systems remains substantially absent from the literature. The extant international algorithmic fairness literature concentrated in the United States context, with seminal contributions examining bias in criminal justice risk assessment (Angwin et al., 2016; Dressel & Farid, 2018), child welfare screening (Eubanks, 2018), and credit scoring (Mehrabi et al., 2021) provides theoretical frameworks and computational methods that are directly applicable to the German context but requires significant adaptation to the specific legal, institutional, and socio-demographic configuration of German public administration. The present study undertakes this adaptation, developing a Socially Grounded Algorithmic Audit Framework (SGAAF) calibrated to the German legal and institutional context and validated through application to ML models representing the specific administrative decision-making domains of SGB II benefit allocation scoring and social housing prioritisation.

The study makes three principal contributions. First, it provides an empirically grounded, computationally rigorous assessment of algorithmic fairness in ML models representative of German public administration functions, generating quantitative evidence on the magnitude and distributional patterns of bias that the existing literature has not previously made available for the German context. Second, it develops and validates the Administrative Justice Ratio (AJR) a novel composite fairness metric designed to assess the alignment between ML model recommendation distributions and the distribution of legally established entitlements under the SGB II framework as a legally anchored fairness measure that bridges the gap between statistical fairness metrics and administrative law compliance assessment. Third, it contributes the SGAAF as a replicable, legally anchored, and sociologically informed audit protocol that German public authorities, AI system providers, and regulatory bodies can deploy to fulfil the fundamental rights impact assessment obligations imposed by the EU AI Act's high-risk system requirements.

2. METHODOLOGY

The Computational Fairness Audit methodology employed in this study integrates three analytical strata: the legal-normative stratum, which establishes the protected characteristics, relevant fairness obligations, and applicable legal compliance thresholds derived from the AGG, the EU AI Act, the GDPR, and the SGB II administrative law framework; the sociological-theoretical stratum, which specifies the fairness concepts and

social justice principles that translate legal non-discrimination obligations into operationalisable statistical fairness metrics; and the computational-statistical stratum, which implements the fairness metrics as quantitative measures applied to ML model outputs and performs the statistical testing required to establish the significance and magnitude of identified bias patterns. The sequential prioritisation of legal and sociological framework specification before computational implementation, explicitly departing from the conventional practice of selecting fairness metrics based on computational convenience or technical familiarity, constitutes the SGAAF's defining methodological commitment. It is justified by the argument that fairness metrics without a theoretically grounded social-justice rationale are analytically arbitrary and legally insufficient for an EU AI Act compliance assessment.

2.1. Dataset Construction: Simulation and Pseudo-Anonymisation Protocols

The study employs two dataset types, each serving a distinct analytical purpose and each constructed in strict compliance with the Bundesdatenschutzgesetz (BDSG), the GDPR, and the specific data minimisation and purpose limitation obligations applicable to the processing of sensitive social data under Artikel 9 DSGVO (GDPR Article 9). The first dataset is a synthetic simulation dataset constructed using a Bayesian generative modelling procedure calibrated against publicly available aggregate statistical distributions from the Bundesagentur für Arbeit's (Federal Employment Agency) published SGB II caseload statistics (Bundesagentur für Arbeit, 2023) and the Statistisches Bundesamt's (Federal Statistical Office) microcensus income and employment distributions. The simulation procedure generates a dataset of 50,000 synthetic SGB II application records, each characterised by 23 socio-demographic and case-specific attributes including age, household composition, employment status, educational qualification level, duration of unemployment, prior benefit receipt history, and a set of geographic and linguistic proxy variables that function as indirect indicators of ethnic or national background without encoding protected characteristics directly. Crucially, the simulation procedure introduces, through calibrated data generation parameters, a systematic variation in historical approval rates across ethnic background proxy subgroups that reflects the historically documented differential approval patterns in SGB II administration specifically, the statistically significant lower approval rates for applicants with Turkish, Middle Eastern, and Sub-Saharan African name-encoded backgrounds documented in correspondence testing studies conducted by the Antidiskriminierungsstelle des Bundes (Federal Anti-Discrimination Agency, ADS) (ADS, 2022). This calibration ensures that the synthetic dataset encodes the structural bias patterns present in real administrative training data without requiring the processing of actual personal data.

The second dataset is a pseudo-anonymised historical record set, compiled under a formal data access agreement with a Bavarian municipal administration, comprising 8,200 de-identified records of municipal social housing (Sozialwohnung) allocation decisions from 2018-2022. Pseudo-anonymisation was implemented through a combination of direct identifier removal, quasi-identifier generalisation (postal code aggregation to district level, age banding to five-year intervals, surname replacement with synthetic surname codes preserving phonological ethnic-origin encoding), and statistical disclosure control through k-anonymity verification with k=5 as the minimum anonymity threshold (Sweeney, 2002). The data access agreement requires that all analytical outputs derived from the pseudo-anonymised dataset be subjected to statistical disclosure risk assessment before publication, and that any results that could enable re-identification of individual records be suppressed. The two-dataset design enables cross-validation of findings: bias patterns identified in the synthetic simulation dataset are assessed for consistency with patterns in the pseudo-anonymised real-world dataset, providing a degree of empirical corroboration for the simulation's calibration validity without requiring the primary analysis to rely on actual sensitive personal data.

2.2. Fairness Metric Specification and Social Justice Grounding

Six fairness metrics were specified before computational implementation, each grounded in a distinct social justice principle and linked to a specific legal compliance obligation. The pre-specification protocol follows the methodological recommendation of Barocas et al. (2019) that fairness metrics be selected based on the normative theory of justice most applicable to the specific decision context being audited, rather than on computational tractability or the most commonly used metric in the existing literature. This is a methodologically significant departure from most published fairness audits, which select metrics post-hoc based on what can be conveniently computed from the available data. The six metrics and their respective social justice groundings are as follows. Demographic Parity (DP) requiring that the proportion of positive

predictions (benefit approvals or housing priority scores above a threshold) be equal across protected group and majority group populations is grounded in the formal equality principle embedded in the AGG's prohibition of direct discrimination and corresponds to the Gleichbehandlungsgrundsatz (equal treatment principle) of German administrative law. Equalised Odds (EO) requiring that both the true positive rate and false positive rate be equal across protected groups is grounded in a procedural fairness principle that holds that individuals with equivalent factual entitlement should have equivalent probability of receiving a positive decision, regardless of group membership, aligning with the SGB II's legal mandate that benefit awards be determined by individual entitlement criteria rather than demographic characteristics.

Calibration Within Groups (CWG) requiring that the model's predicted probability of a positive outcome be equally accurate across subgroups, such that a predicted probability of 0.7 corresponds to a 70% actual positive rate equally for all groups is grounded in the epistemic justice principle that the model's uncertainty representations should be equally reliable across all population groups rather than systematically overconfident for majority group predictions and underconfident for minority group predictions. Individual Fairness (IF) requires that individuals with similar relevant characteristics receive similar predictions and is grounded in Dworkin's (1977) principle of treating individuals as individuals rather than as representatives of group averages, and is operationalised through a distance metric comparing feature vectors across the individual cases in the test dataset. Counterfactual Fairness (CF) requiring that an individual would receive the same prediction in a counterfactual world where their protected characteristic was different while all causally relevant non-protected characteristics remained constant is grounded in a causal justice principle that directly addresses the mechanism of proxy discrimination, where a protected characteristic is encoded in the prediction through causally downstream variables (Kusner et al., 2017). The novel Administrative Justice Ratio (AJR) introduced by this study measures the ratio of the ML model's recommendation distribution to the distribution of legally established entitlements under the SGB II framework, assessing the degree to which the model's predictions diverge from what a legally compliant, entitlement-based decision process would produce for each demographic subgroup. The AJR is computed as the ratio of the subgroup's model-recommended approval rate to the subgroup's entitlement-calculated approval rate, with a value of 1.0 indicating full alignment and values below 1.0 indicating that the model systematically under-recommends relative to legal entitlement for that subgroup.

2.3. ML Model Architectures and Training Protocol

Three ML model architectures were trained on the synthetic simulation dataset: logistic regression (LR), gradient boosting (GB), and a fully connected deep neural network (DNN), representing the spectrum of computational complexity and interpretability that characterises real-world administrative ML deployments in Germany. Logistic regression represents the most transparent and interpretable end of this spectrum; it has been deployed in several German administrative contexts precisely because its decision logic can be expressed in a mathematically readable coefficient form that supports regulatory scrutiny. Gradient boosting, implemented using the XGBoost library (Chen & Guestrin, 2016), is a class of ensemble tree-based models that achieve the highest predictive performance on structured tabular data and have been increasingly adopted in scoring and risk assessment applications across German financial and social service contexts. The deep neural network, implemented using PyTorch with three hidden layers of 256, 128, and 64 neurons with ReLU activation functions and dropout regularisation, represents the most computationally opaque architecture and serves as the audit's highest-complexity test case. All three models were trained using an 80/20 stratified train-test split, with hyperparameter optimisation conducted through five-fold cross-validation on the training set. Standard classification performance metrics accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC-ROC) were computed for each model on the held-out test set, followed by the six fairness metric computations disaggregated by the four protected subgroup definitions: Turkish-German background, refugee or asylum-seeker background, single-parent household, and disability status. The pseudo-anonymised social housing dataset was used for a parallel fairness audit of a gradient-boosting allocation scoring model, with results reported as a cross-domain robustness check of the primary simulation dataset findings.

3. RESULTS AND DISCUSSION

3.1. Classification Performance vs. Fairness Performance: The Accuracy-Equity Trade-off

The computational audit results reveal a pattern of systematic, statistically significant tension between classification accuracy and fairness metric performance, which constitutes the study's most practically consequential finding for German public administration AI procurement and governance. Across all three model architectures, the gradient boosting and deep neural network models achieve substantially higher classification accuracy than logistic regression on the standard performance metrics: the GB model achieves an AUC-ROC of 0.847 and an overall accuracy of 79.3% on the test set, the DNN achieves an AUC-ROC of 0.861 and an accuracy of 81.1%. In contrast, logistic regression achieves an AUC-ROC of 0.791 and an accuracy of 74.6%. These performance differentials are consistent with the well-established empirical pattern that more complex model architectures provide superior predictive performance on structured administrative datasets. In conventional AI procurement practice, these results would lead directly to selecting the DNN or GB model as the preferred deployment architecture.

The fairness metric analysis fundamentally complicates this procurement logic. The DNN, which achieves the highest classification accuracy, also exhibits the most severe fairness violations across the protected subgroups. The Demographic Parity gap, computed as the absolute difference in positive prediction rates between the majority population reference group and each protected subgroup, is 0.187 for the DNN, 0.164 for the GB model, and 0.092 for the logistic regression model. In substantive terms, this means that the DNN generates a benefit approval recommendation rate approximately 18.7 percentage points lower for individuals in the Turkish-German background subgroup than for equivalent individuals in the majority reference group, after controlling for the legally relevant eligibility criteria encoded in the model features. The logistic regression model's demographic parity gap of 9.2 percentage points represents a substantially lower, though still legally significant, deviation from the Gleichbehandlungsgrundsatz. The refugee or asylum-seeker background subgroup exhibits the largest demographic parity gaps of any protected group across all three models: 0.213 for the DNN, 0.189 for the GB model, and 0.114 for logistic regression, reflecting the severity of the proxy discrimination encoded in the historical training data distributions for this population subgroup.

The Equalised Odds analysis reveals a specific pattern of false negative rate disparities the rate at which eligible individuals from protected subgroups are incorrectly recommended against benefit approval that has direct legal consequences under the SGB II entitlement framework. The DNN model's false negative rate for the Turkish-German background subgroup is 0.341, compared to a majority group false negative rate of 0.198, representing a differential of 14.3 percentage points that means that eligible Turkish-German applicants are incorrectly denied a positive recommendation at a rate 72% higher than equivalent majority group applicants. The gradient boosting model exhibits a similar differential of 11.8 percentage points. These false negative rate disparities represent, in administrative law terms, a systematic failure of the Rechtmäßigkeitsprinzip (principle of lawfulness) the requirement that administrative decisions conform to applicable law because the SGB II establishes entitlement to benefit as a legal right conditional on objective eligibility criteria, and a model that systematically underestimates entitlement for a specific ethnic background subgroup generates recommendations that, if acted upon, would constitute unlawful denial of legally established rights.

3.2. The Administrative Justice Ratio: Legal Entitlement vs. Model Recommendation

The Administrative Justice Ratio provides the most legally tractable expression of the audit findings, translating the statistical fairness metric violations into a direct comparison between model-recommended distributions and legally established entitlement distributions, interpretable by administrative law practitioners without specialist statistical training. The AJR values computed for the simulation dataset across the four protected subgroups reveal a consistent pattern of model under-recommendation relative to legal entitlement, most severe for the refugee and asylum-seeker background subgroup. The DNN model generates an AJR of 0.71 for this subgroup, meaning it recommends positive decisions for only 71% of the subgroup's members who are legally entitled to a positive decision under the SGB II criteria, compared to a majority-group AJR of 0.93 for the same model. The gradient boosting model's AJR for the refugee background subgroup is 0.74, while logistic regression achieves an AJR of 0.82. For the Turkish-German background subgroup, the DNN's AJR is 0.76, the GB model's is 0.79, and logistic regression's is 0.85. The single-parent household subgroup shows more moderate AJR disparities: DNN 0.83, GB 0.86, LR 0.91. In contrast, the disability status

subgroup exhibits the smallest AJR gap of the four protected groups, DNN 0.88, suggesting that the disability-related features in the simulation dataset are less susceptible to the proxy discrimination mechanisms that most severely affect ethnically-encoded subgroups.

The AJR findings have direct implications for the EU AI Act conformity assessment obligations applicable to these model architectures as high-risk AI systems in public administration contexts. Article 9 of the AI Act requires providers of high-risk AI systems to implement a risk management system that identifies and addresses the risks to fundamental rights associated with the system throughout its lifecycle. The AJR disparities documented in this study, which indicate that the models systematically generate recommendation distributions that diverge from legally established entitlement distributions in ways that disadvantage specific ethnic and socio-demographic groups, constitute precisely the fundamental rights risks that the Act's risk management obligation is designed to identify and mitigate. A fundamental rights impact assessment conducted on any of the three model architectures using the SGAAF audit protocol would be required to disclose these disparities, and the DNN and GB models would be assessed as failing to meet the fundamental rights protection standard required for lawful deployment in SGB II administration under the Act's high-risk system requirements.

3.3. Counterfactual Fairness and Proxy Discrimination: The Mechanism of Ethnic Bias Encoding

The Counterfactual Fairness analysis provides the most mechanistically illuminating findings of the audit, identifying the specific causal pathways through which ethnic background information, which is explicitly excluded from the model feature set as a protected characteristic, is nonetheless encoded in the model's predictions via causally downstream proxy variables. The counterfactual fairness scores computed through a structural causal model that estimates the probability that an individual's prediction would change if their ethnic background were counterfactually altered while holding causally non-descendant variables constant reveal substantial counterfactual unfairness in both the GB and DNN models: 34.7% of Turkish-German background individuals in the test set would receive a different (positive) prediction if their background were counterfactually changed to majority-German background, while holding their objective eligibility characteristics constant. For logistic regression, the counterfactual unfairness rate is 21.3%, substantially lower but still legally significant.

Decomposition of the counterfactual effects using SHAP (Shapley Additive exPlanations) values identifies three proxy-variable clusters that account for the majority of the counterfactual fairness violations. The first is postal code encoding: applicants from postal code areas with above-median concentrations of Turkish-German residents receive systematically lower benefit scoring contributions from the postal code feature, reflecting the historical correlation between residential segregation patterns and benefit approval rates in the training data. The second is employment sector proxy: the feature encoding prior employment sector encodes a statistical association between sectors with high Turkish-German employment concentrations particularly manufacturing, gastronomy, and cleaning services and lower benefit approval rates, reflecting historical administrative patterns rather than legally relevant eligibility criteria. The third is surname phonological encoding: while surnames are not directly included as model features, the synthetic data generation procedure introduced a surname-encoded proxy variable representing the phonological plausibility of a surname's ethnic origin that the GB and DNN models learned to use as a predictive signal through its statistical association with the outcome variable in the training data. The identification of these specific proxy pathways is not merely an academic contribution; it provides the actionable specification of feature re-engineering interventions postal code de-sensitisation through aggregate substitution, sector category re-coding to suppress ethnically encoded correlations, and surname proxy variable removal that can reduce counterfactual unfairness to legally tolerable levels while maintaining adequate model performance.

3.4. Social Housing Dataset Cross-Validation: Robustness of Findings

The parallel fairness audit of the gradient-boosting social housing allocation scoring model trained on the pseudo-anonymised Bavarian municipal dataset provides cross-domain robustness validation of the primary simulation dataset findings. The social housing audit context presents a structurally different allocation mechanism: housing priority scoring is not a binary entitlement determination but a ranked prioritisation among eligible applicants for a constrained supply of housing units, requiring adaptation of the fairness

metrics to a ranked outcome context. The Demographic Parity metric measures the proportion of each protected subgroup represented among the top quartile of priority scores, and the AJR measures the ratio of a subgroup's representation in top-quartile allocations to its representation in the legally eligible applicant pool. The social housing audit results reveal demographic parity disparities that are broadly consistent with the SGB II simulation findings: Turkish-German background applicants are under-represented in the top quartile of housing priority scores by 12.3 percentage points relative to their share of the legally eligible applicant pool, and refugee-background applicants by 16.8 percentage points. Single-parent household applicants show a reversed pattern: moderate over-representation in top-quartile scores of 4.2 percentage points, suggesting that the housing allocation model has learned to partially compensate for the structural housing disadvantage of single-parent households through a feature interaction that does not appear in the benefit allocation model. These cross-domain consistencies strengthen the empirical case for the generalisability of the SGAAF audit findings beyond any single model or administrative domain.

3.5. The Socially Grounded Algorithmic Audit Framework: Design and Regulatory Application

The Socially Grounded Algorithmic Audit Framework synthesises the methodological approach and empirical findings of this study into a replicable protocol architecture for algorithmic fairness auditing in German public administration. The SGAAF is structured as a five-phase protocol, with each phase producing a documented output that contributes to the fundamental rights impact assessment required by the EU AI Act for high-risk AI systems. Phase 1, Legal-Normative Framework Specification, requires auditors to identify the applicable protected characteristics under the AGG and relevant sector-specific law, the specific fairness obligations imposed by the administrative law framework governing the decision domain, and the legal compliance thresholds against which fairness metric violations will be assessed. The outputs of this phase are a Rechtliche Fairness-Spezifikation (Legal Fairness Specification) document and a Schutzgruppen-Taxonomie (Protected Group Taxonomy) that define the audit's legal boundary conditions.

Phase 2, Sociological Justice Framework Specification, requires auditors to articulate the social justice principles most applicable to the specific decision domain selecting from the range of formal equality, procedural fairness, distributive justice, and epistemic justice principles that the fairness metrics operationalise and to justify the selection of specific fairness metrics from this theoretical framework rather than selecting metrics based on computational convenience. The output is a Soziale Fairness-Begründung (Social Fairness Justification) document that links each selected fairness metric to its applicable social justice principle and legal compliance obligation, establishing the normative foundation for the subsequent computational analysis. Phase 3, Data Governance and Privacy-Compliant Dataset Construction, operationalises the data access, pseudo-anonymisation, and synthetic data generation protocols required to construct audit datasets that are adequate for fairness assessment while complying with the BDSG, GDPR, and sector-specific data protection obligations. The output is a Datenschutz-Konformitätsbericht (Data Protection Compliance Report) certifying the audit dataset's compliance with applicable data protection law.

Phase 4, Computational Fairness Audit Execution, implements the prespecified fairness metrics against the ML models under evaluation, including all six metrics specified in Section 2.2, plus the Administrative Justice Ratio for administrative law compliance contexts, and produces disaggregated fairness metric results with bootstrapped confidence intervals for statistical significance assessment. The output is a Fairness-Audit-Bericht (Fairness Audit Report) presenting the full metric results, protected subgroup comparisons, counterfactual fairness decomposition, and SHAP-based proxy discrimination pathway analysis. Phase 5, Regulatory Compliance Assessment and Remediation Specification, translates the Fairness-Audit-Bericht findings into a legal compliance assessment against the EU AI Act's fundamental rights impact assessment requirements, identifying specific metric violations that constitute high-risk findings requiring remediation before deployment authorisation. Where violations are identified, the phase produces a Abhilfemaßnahmen-Spezifikation (Remediation Specification) documenting the feature re-engineering, model retraining, or deployment constraint interventions required to bring the model into compliance with the applicable fairness obligations. This five-phase architecture positions the SGAAF as a legally defensible, technically rigorous, and sociologically informed audit instrument that German public authorities and AI system providers can deploy to meet the EU AI Act's conformity assessment obligations for high-risk administrative AI systems.

3.6. Implications for German AI Governance Policy and the EU AI Act Implementation

The empirical findings of this audit carry direct and urgent implications for the implementation of the EU AI Act's high-risk system requirements in the German public administration context. The finding that the most technically performant models, as measured by conventional classification accuracy metrics, exhibit the most severe fairness violations fundamentally challenges the technical sufficiency standard that has implicitly governed German public-sector AI procurement: the assumption that procuring a technically validated, high-accuracy ML system constitutes adequate AI governance. The SGAAF audit demonstrates empirically that technical accuracy validation and fundamental rights compliance assessment are not merely different aspects of the same evaluation but are potentially in systematic tension with each other, and that AI Act compliance requires the latter to take priority over the former in selecting systems for high-risk public administration deployment.

The German Federal Government's AI Action Plan and the Bundesnetzagentur's (Federal Network Agency) emerging role as the national market surveillance authority under the AI Act will require the development of standardised conformity assessment procedures for high-risk public administration AI systems that are technically adequate for detecting the proxy discrimination mechanisms identified in this study. The SGAAF provides a methodological foundation for the development of such standardised procedures. The study recommends its adoption with iterative refinement through a community of practice, including the Antidiskriminierungsstelle des Bundes, the Bundesbeauftragte für Datenschutz und Informationsfreiheit (Federal Commissioner for Data Protection and Freedom of Information), and independent algorithmic audit specialists as the basis for a national standard for fundamental rights impact assessment of high-risk administrative AI systems. The study further recommends the mandatory pre-deployment disclosure of SGAAF audit results for any ML system deployed in SGB II administration, social housing allocation, or child welfare risk assessment contexts, with results provided to affected benefit recipients as part of the algorithmic transparency obligations arising from the GDPR's Article 22 right to explanation and the AI Act's Article 13 transparency obligations.

4. CONCLUSION

This study has provided the first computationally rigorous, legally anchored, and sociologically grounded empirical assessment of algorithmic fairness in ML systems representative of German public administration functions, generating evidence that confirms the presence of statistically significant and legally consequential bias against Turkish-German, refugee-background, and other protected population subgroups across all three tested model architectures. The inverse relationship between classification accuracy and fairness metric performance with the most accurate models exhibiting the most severe demographic parity gaps and equalised odds violations constitutes a finding that fundamentally challenges the technical sufficiency assumptions embedded in current German public sector AI procurement practice and demands a regulatory response that prioritises fundamental rights compliance assessment over performance optimisation in high-risk administrative AI deployment decisions.

The Administrative Justice Ratio introduced by this study provides a legally tractable fairness measure that translates statistical audit findings into administrative law compliance language accessible to public administrators, legal practitioners, and regulatory authorities, bridging a critical communication gap between the technical algorithmic fairness literature and the practical governance requirements of the EU AI Act's high-risk system framework. The counterfactual fairness and SHAP decomposition analysis identifies the specific proxy discrimination pathways postal code encoding, employment sector proxy variables, and surname phonological encoding through which ethnic background information is encoded in model predictions despite its explicit exclusion from the feature set, providing actionable remediation specifications that can guide the technical corrections required for legal compliance.

The Socially Grounded Algorithmic Audit Framework represents the study's most enduring methodological contribution: a replicable, legally anchored, five-phase audit protocol that integrates legal-normative specification, sociological justice theory, privacy-compliant data governance, computational fairness measurement, and regulatory compliance assessment into a coherent instrument for the pre-deployment and in-deployment evaluation of high-risk administrative AI systems. As the EU AI Act's phased application schedule brings its full requirements into force by 2027, the SGAAF provides a technically validated and

legally grounded foundation for the development of the national conformity assessment procedures that German public authorities, AI system providers, and regulatory bodies must establish to fulfil their obligations under Europe's most ambitious and consequential AI governance framework.

REFERENCES

Antidiskriminierungsstelle des Bundes (ADS). (2022). Diskriminierung im Bereich des Sozialrechts: Ergebnisse des Monitorings 2020-2022. ADS.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica.

Barocas, S., & Hardt, M. (2017). Fairness in machine learning—proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS).

Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning: Limitations and opportunities. fairmlbook.org.

Bundesagentur für Arbeit. (2023). Statistik der Grundsicherung für Arbeitsuchende nach dem SGB II: Jahresbericht 2022. Bundesagentur für Arbeit.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. https://doi.org/10.1145/2939672.2939785

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153-163. https://doi.org/10.1089/big.2016.0047

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science Advances, 4(1). https://doi.org/10.1126/sciadv.aao5580

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 214-226. https://doi.org/10.1145/2090236.2090255

Dworkin, R. (1977). Taking rights seriously. Harvard University Press.

Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor—St Martin's Press.

European Parliament. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689.

Kuhlmann, S., & Heuberger, M. (2021). Digitalisation of public administration in Germany: Characteristics, challenges and outlook. Public Management Review, 23(1), 1-25. https://doi.org/10.1080/14719037.2021.1872916

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. Advances in Neural Information Processing Systems, 30, 4066-4076.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1-35. https://doi.org/10.1145/3457607

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), 557-570. https://doi.org/10.1142/S0218488502001648

