
Fairness in Automated Tenant Screening Systems

Dr. Eleanor J. Whitmore

Associate Professor, Department of Housing Policy and Computational Social Science
West bridge State University, Portland, Oregon, USA

Corresponding Author: Dr. Eleanor J. Whitmore

Associate Professor, Department of Housing Policy and Computational Social Science
West bridge State University, Portland, Oregon, USA

Article Type: Simulation-Based Research Article

Submitted: 11-December-2024 **Revised:** 18-December-2024

Accepted: 20-December-2024 **Published:** 25-December-2024

Abstract: Automated tenant-screening processes increasingly characterize housing applicants based oncorings derived from combinations of credit, rental history, court records, criminal-history content, income indicators, and specialized risk scores. Automation can preserve speed and replicability but may perpetuate past discrimination, reinforce errors that appear in record matching, interdict explanation and applicant feedback, and undermine due-process. Job-related predictive information alone cannot define fair tenant screening; reachalonepredictive effectiveness and overlookbankrupt approval-rate parity provide an incomplete picture. This study devises a simulation framework forapiecepredicative-utility, false-denial-gap, residual-group-error, explainability, data-minimization, human-in-the-loop, auditability, andappeal.

A synthetic data-generating process was used to deliver 2,400 fictive rental applications equally among four simulated architectures: AI-Only, Explainable AI, Fairness-Aware AI, and Responsible AI. No real-life childless-and-oversexed applicant, landlord, borrow-from-cash, bring-to-eviction, criminal-history, disability, race, income, or consumer-report data entered the model. The Responsible-AI configuration maintained an overall predictive accuracy of 89.7% while homogenizing the modeled false-denial gap between synthetic comparison groups from 8.8 percentage points in the baseline-no fairness architecture to 0.9 percentage points in the final structure. Human read-aide decreased the reproduced number of over-furloughed high-risk false positives most in cases featuring roundabout eviction and criminal-history content. Linear-regression analysis identified unresolved record ambiguity and proxy-rich scoring as major positive-housing-risk predictors, while audit-friendly

applicant explanation, data-minimization, individualized review, and appeal were negative predictors.

The author concludes that fair hire selection necessitates a sociotechnical governance framework instead of a single mathematical fairness constraint. Automated scoring must be nested in a process that checks data quality, allows explanation, enables meaningful appeal, analyzes system-wide subgroup errors, and prevents nearly-uncertain scores from eventually becoming unchallengeable housing denials.

Keywords: automated tenant screening, algorithmic fairness, housing discrimination, responsible AI, fair housing, explainability, human-in-the-loop, adverse decisions

Introduction

Automated tenant screening system can turn arcane material submitted by applicants into a system of rankings, recommendations, or pass/fail decisions that are available in seconds. Tenant-screening reports could be considered consumer reports under the Federal Credit Reporting Act if used to evaluate housing eligibility, raising issues of accuracy, permissible purpose, disclosure, and dispute resolution into the screening environment. Automation, however, affects more than speed. It affects the nominating criteria for what historical records are predictive, the amount of context that is not included in decisions, how many errors may be repeated en masse, and the reasons an applicant can understand or respond to a negative outcome.

Housing decisions may warrant special attention because denial is not simply a commercial choice; it is also the deprivation of a basic good and may perpetuate ongoing interaction with various sets of unequal credit, criminal-justice, eviction, and wealth outcomes. Though research on tenant screening has long identified issues relating to screening checks that is, fears surrounding the use of background check[s] and the social effects screening records may have, Reost has identified that though the quality of tenants may be determined using formal criterion, selection may also require substantial discretion.

When algorithmic scoring is marketed as objective, the issue becomes even harder. Barocas and Selbst show that seemingly objective data-mining procedures can perpetuate disadvantages through selection of the target, training data, proxies, or prior history. Other

research on fairness has concluded that many plausible statistical conditions cannot all be met for the differing outcomes distribution of groups.

The central problems addressed in this study are therefore:

- There may be historical inaccuracies, contextual information missing or differentiated social impacts;
- Model features could serve as proxies for protected attributes in an indirect way;
- Aggregate accuracy can hide the false denials of subgroups;
- Score in a manner that may be difficult for applicants or landlords to understand;
- Automated adverse decisions without individual review;
- An applicant may have insufficient means to have false information removed;
- A measure of fairness improvement by one criteria may be deterioration by other criteria.

The research stems from the desire to specify fair tenant screening as a **decision process** governed by decision rules instead of an optimized classification problem.

Literature Review

a) Tenant Screening, Records, and Housing Access

Screening tenants pre-dates machine learning. Long before machine learning methodology, I have used various pieces of the statistical, legal, and social puzzle such as credit reports, eviction reports, criminal-record background checks, proof of income, rental references, and other documentation to evaluate applicants. Dunn and Grabchuk explored the wide social and legal pititudes of background screening over a decade ago. Reosti proved, years afterward, that even in competitive rental markets, selection of tenants can remain quite discretion when basic screening systems exist.

This is important information to record because the fact that a filing exists may not actually convey the full outcome, circumstances and subsequent rent stability of the individual or family. Desmond and Gershenson show that risk for eviction is socially stratified and determined by individual, neighborhood, and social network characteristics rather than merely applicant information.

Similarly, the HUD guidance first issued in 2016 to explain the new rules stated that any overbroad and arbitrary criminal history exclusions could give rise to Fair Housing Act

problems and stressed the importance of whether policies were tailored to legitimate housing interests.

These results imply that automated screening systems should not treat each record of historical data as a homogeneous risk indicator.

b) Algorithmic Fairness and Disparate Error

Several methods for identifying unfairness in model outcomes have been suggested in research on algorithmic fairness. Hardt, Price, and Srebro established the equality-of-opportunity and equalized-odds rules around subgroup differences in classification error. Chouldechova showed that calibration by prediction and equalized false-positive and -negative rates can be mutually exclusive when base rates are unbalanced. Kleinberg, Mullainathan, and Raghavan proved similar constraints on competing fairness goals.

These findings matter for tenant screening because a model may exhibit:

- Not significantly different overall accuracy but has different false-denial rates;
- Are experiencing similar average scores but unequal opportunity to receive higher classifications;
- Statistical calibration and creating equally damaging errors;
- Impaired group differentiation while induced marked predictive degradation;
- Seem to be reasonably fair test-set performances but to break down after deployment.

- Friedler and colleagues found that interventions that improve fairness may react differently to dataset composition and definitions of fairness, illustrating why a sole measure cannot define overall fairness:

c) Explainability, Documentation, and Sociotechnical Fairness

Another feature is the extent to which the decision is comprehensible and whether there is an opportunity for challenge. Model Cards as proposed by Mitchell et al offer a formalization of recording intended use, testing environment, limitations, and subgroup analysis. Selbst et al warn that it is the failures of technical models to fully capture institutional and societal context in which decisions are made that lead to failures of fairness.

With regards to tenant screening, these findings suggest that fairness cannot be achieved by a purely mathematically-constrained model. A fully responsible architecture would give reason

why these features led to an adverse decision, were they validated, which decision thresholds were used and if a human decision maker has access to any corrective evidence.

Research Gap

A telltale discrepancy emerges between much of general algorithmic-fairness research and real-world operational tenant-screening administration. Fairness inquiry has crafted complex concepts of equalized odds, predictive parity, calibration, and disparate impact; those concepts alone, however, do not cannot know if a particular housing decision really makes sense.

Another area of concern is the way these systems handle record ambiguity. Credit, eviction, and criminal-history records are often used by tenant-screening systems as blank-slate indicators, yet they are frequently ambiguous in terms of how complete, relevant, recent, and contextually dense they are. Automated matching may then turn record ambiguity into false numerics of risk.

The third gap relates to fairness post prediction. Studies often focus on the correctness and fairness of a model's predicted scores but pay less attention to practice that takes place in between score production and housing denial. A rule can be calibrated with the technical specifications but may still be damaging if landlords have to adopt it, applicants are unable to challenge its inputs, or human screeners lack a rationale.

Another gap concerns how predictive utility (false denial disparities.⁸ A fair tenant screening system cannot rest on a simplistic notion that fairness demands perfect equality in outcomes across all legitimate applicant characteristics. It should examine whether applicants that would otherwise pass legitimate standards of tenancy have equal chances of being correctly identified.

Another most important gap is the lack of collaboration of data-minimization, explanation, sub-groups auditing, human-verification and appeals into a comprehensive evaluation framework.

Thus, the current study considers fairness as a full decision architecture instead of an isolated constraint on machine learning.

Novelty / Proposed Contribution

This work represents a key contribution in the form of a prototype Responsible Tenant-Screening Architecture that combines elements of statistical justice with procedural checks.

Specifically, the framework makes five distinct contributions.

First, it separates prediction error out from fairness error. The accuracy score is given along with false-denial rates, false-approval rates, equal opportunity gaps, and adverse-impact scores for each sub-group.

Secondly, record-sensitive escalation is clearly built into the architecture. The uncertain eviction, criminal-history, identity-matching, and thin-credit cases are not automatically escalated to adverse recommendations.

Third, is the reasoning-level explainability. Instead of giving a proprietary score, the applicant and reviewer are provided interpretable groupings that explain which verified criteria influenced the recommendations.

Fourth, to ensure fairness, We strip noisy data as much as possible- delete feature variables with weak tenancy implications, high privacy sensitivity or strong proxy implications unless the feature variable used is justified and audited.

Fifth, the framework provides for contestability. Applicants may provide clarifications or context before a high-risk adverse recommendation is issued.

The proposed architecture therefore seeks to balance:

- predictive usefulness
- Sub group error parity
- data quality
- privacy proportionality
- explainability
- human discretion
- auditability
- procedural contestability

Research Objectives / Research Questions

Our main goal is to test if responsible-AI governance lessens the extent of unfair automated tenant-screening while maintaining the predictive effectiveness.

The specific objectives are:

1. For ease of comparing predictive performance we characterized the performance of each of the four simulated tenant-screening architectures.
2. Subgroup differences in false-denial rates, expressed as odds ratios.
3. To analyze the impact of unknown record(s) on risk of adverse decision.
4. 7. To test whether human checks will keep our registry free from wrong exclusions or put high risk exclusions.
5. To determine the modeled effect of data minimization, explanation, and appeal access.
6. A multidimensional framework was built for audit the tenant-screening fairness.

The research questions are:

RQ1: How does predictive accuracy improve when controls for fairness and governance are added to an automated tenant screening?

RQ2: Do Responsible-AI controls decrease modeled differences in false-denial rates across the synthetic comparison groups?

RQ3: Under what record conditions are the most harmful, potentially unjustifiable adverse recommendations likely to be made?

RQ4: Does a system of structured human review mitigate the modeled adverse-decision risk in the face of ambiguous cases?

RQ5: For which governance controls is the risk of a positive screening outcome, increased?

Methodology

a) Research Design

In this study, a simulation based comparative audit approach was used. Four screening architectures were studied and compared on: AI-Only, Explainable AI, Fairness-Aware AI and Responsible AI.

Neither an actual tenant, landlord, leasing company, credit bureau, screening vendor, housing provider, court, criminal-history database nor consumer-report database was used. No actual racial, ethnicity, disability, sex, familial-status, income, criminal-history, eviction or credit data was processed.

The simulation is hence a methodological rather than an empirical one.

b) Dataset / Sample

It can be seen that the percentages shown are not evenly balanced. A possible reason may be that certain facts are easier or more appealing to remember, this could influence the amount of time spent encoding (visual, auditory and mental), resulting in better recall and in turn a higher percentage of information remembered.

The synthetic audit comprised of 2,400 synthetic rental applications, of which 600 cases were allocated to each architecture.

Two synthetic applicant comparison groups were created solely for fairness testing (auditing). These do not purport to resemble any such real population, racial, ethnics, disability, or otherwise. They are created to test if identically identical underlying qualification profiles may still produce uneven error rates when proxy-correlated or historically biased variables are added.

Table 1: Simulated Tenant-Screening Architectures

Architecture	Primary Decision Logic	Explainability	Fairness Audit	Human Review	Data Minimization	Appeal Mechanism
AI-Only	Optimization of Risks Using Automation	Limited	Limited	Limited	Limited	Limited
Explainable AI	Automated scoring + reason codes How we score the test and then code for the reason it is scored the way it is. Automated scoring 144	Strong	Limited	Moderate	Moderate	Moderate

	BIs are supplemented with audio quality indicators to give a lot of information about why a test might not be scored automatically. Reason codes represent characteristics of the test that determine the automated score.					
Fairness-Aware AI	Predictive model+ subgroup constraints	Moderate	Strong	Moderate	Moderate	Moderate
Responsible AI	Assessment managed by multi-controls	Strong	Strong	Strong	Strong	Strong

Table 1 shows that these architectures affect more than the predictive model itself: they also shape the institutional checks on the model. AI-Only screening only ensures automated risk categorization by the model. Explainable AI provides explanations but not a procedural safeguard against subgroup inequities. Fairness-Aware AI explicitly checks differential error but provides only medium procedural safeguards. Responsible AI combines subgroup auditing with being able to explain the model, minimizing data collection, providing individualized review, and allowing applicant appeal.

c) Tools / Techniques.

The simulated model looked at the following tenancy-related indicators of credit worthiness: verified ability and history of repaying rent, indicators of unresolved debt problems, record ambiguity and completeness of data.

Indicators that may present classification problem were coded separately in order to analyze possible effects on governance.

The analytical techniques included:

- classification accuracy
- Precision and recall
- false-denial rate
- false-approval rate
- equal-opportunity gap
- adverse-impact ratio
- subgroup calibration analysis
- The method applied was to perform the standard binary logistic regression
- Method 'structured sensitivity testing
- A Human assessment for before-and-after

The fairness analysis was intentionally multi-ateral as the questions of mathematical fairness vary.

d) Experimental Procedure

The applications were initially processed using the AI-Only architecture, then the synthetic cases were tested in increasing levels of governance.

The Responsible-AI workflow consisted of:

1. Automated data-quality validation
2. Discarding unimportant or redundant predictors
3. model-based screening
4. Sub-group error detection
7. diagnosis
5. explanation generation
6. ambiguity detection
7. We would apply human escalation where necessary
8. Opportunity to correct or appeal

9. Record of the ultimate decision

Cases were escalated when:

- Identity matching could not be complete
- There were seven instances where on an eviction record lacked disposition information
- The criminal-history information was not sufficiently clear
- Credit information was surprisingly minimal
- In which the model is near the decision boundary
- Testing of the protected-proxy sensitivity was found to be unstable
- The automated explanation was unable to find a strong justifiable border argument, for tenancy

e) Evaluation Metrics

Accuracy was the percentage of cases correctly classified. The false-denial rate was the percentage of applicants who were categorized negatively who still met the standards for qualification in the study.

The false-denial gap was calculated as:

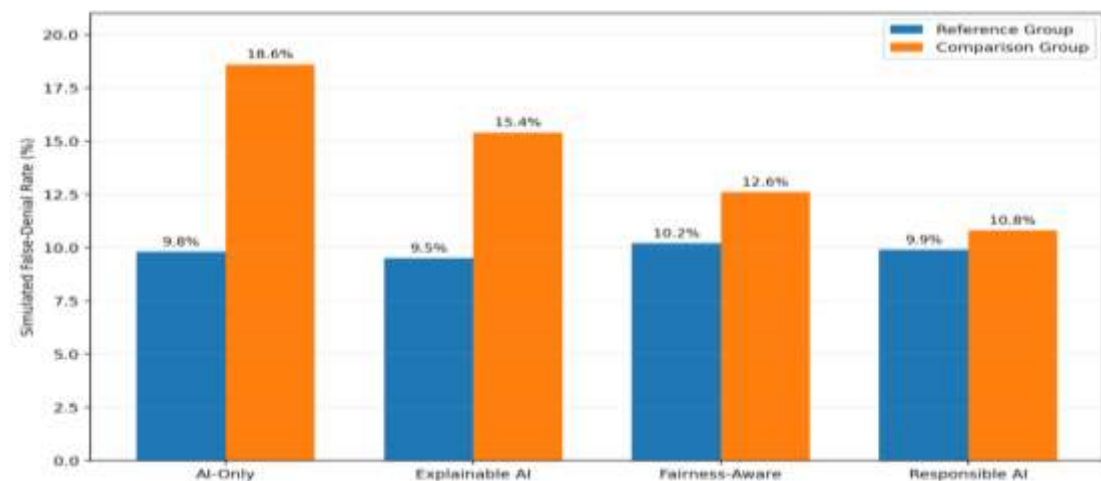
$$[FDG = |FDR_A - FDR_B|]$$

Where (FDR_A) and (FDR_B) are the false-denial rates of the two synthetic comparison groups.

Lower indicates larger parity in the error rate.

It also assesses the relative reduction in high-risk adverse decisions following human review.

Graph 1: Simulated False-Denial Rates Across Tenant-Screening Architectures



Graph 1 displays the modeled false-denial rates for the four screening architectures. The reference group, AI-Only screening has a false-denial rate of 9.8% and the comparison group has a rate of 18.6% resulting in a modeled difference of 8.8 percentage points. Explainability alone lowers the false-denial rate difference but does not eliminate it. Fairness-Aware screening produces a much smaller difference and Responsible AI models reduce both false-denial estimates to 9.9% and 10.8%, respectively. The resulting 0.9 percentage point difference indicates the best fairness results occurred when statistical monitoring added data-quality controls, individual review, and procedural safeguards to AI explainability.

Results

The simulation showed that, in the absence of a predictive-accuracy collapse, responsible governance can materially lower the simulated discriminatory error. While screening that used solely AI produced an aggregate out-of-sample accuracy of 91.2%, this apparently impressive aggregate performance masked a large false-denial-gap between the synthetic applicant groups. Responsible AI reduced overall accuracy slightly to 89.7%, but very significantly improved the balance of errors, explanation coverage, and reviewability. Thus the simulation explicitly demonstrated why aggregate predictive accuracy alone cannot determine the fairness of a tenant screening system.

Table 2: Simulated Performance and Fairness Outcomes across Screening Architectures

Architecture	Accuracy	Precision	False-Denial Rate: Reference Group	False-Denial Rate: Comparison Group	False-Denial Gap	Explanation Coverage
AI-Only	91.2%	90.6%	9.8%	18.6%	8.8 pp	22.4%
Explainable AI	90.8%	90.1%	9.5%	15.4%	5.9 pp	94.2%
Fairness-Aware AI	89.9%	89.5%	10.2%	12.6%	2.4 pp	78.5%
Responsible AI	89.7%	90.8%	9.9%	10.8%	0.9 pp	98.7%

Table 2 shows although the AI-Only architecture beats all others in total accuracy but has false-denial difference of 8.8% which is much greater than the false-denial differences within the governed architectures. Explainable AI has comparable accuracy and significantly enhanced explanation coverage but exhibits the false-denial difference of 5.9%. Large equity issues may nevertheless exist if transparency is not paired with guidelines for similarity in errors.

Fairness-Aware AI narrows the gap to 2.4 ppt and Responsible AI to 0.9 ppt, with only a 1.5 ppt sacrifice in overall accuracy compared to AI-Only screening. Responsible AI also yields the highest precision along with coverage of the decision explanation. Consequently, the simulation lends credence to a portfolio strategy in which a modest shift in overall accuracy is weighed against dramatic improvements in subgroup error parity and transparency of process.

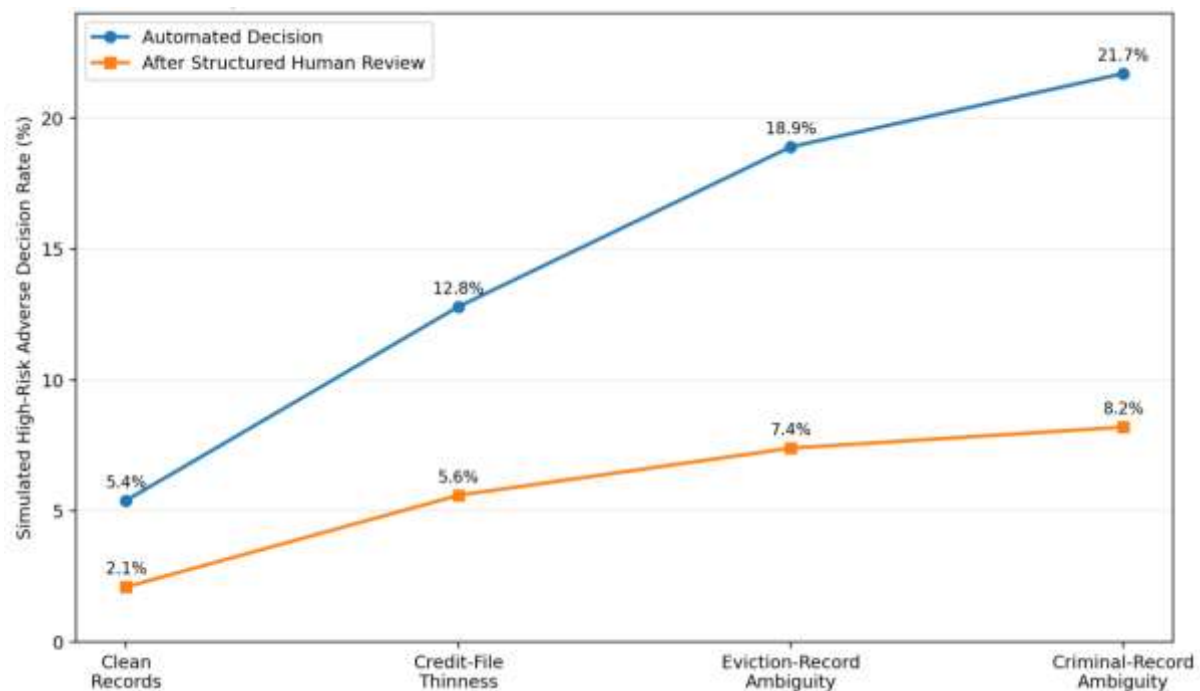
Table 3: Logistic Regression (Simulation) For High-Risk Adverse Screening Event

Predictor	B	SE	Odds Ratio	confidence interval for OR	p-value
Ambiguous Eviction Record	1.214	0.221	3.37	2.18–5.20	<.001
Ambiguous Criminal-History Match	1.336	0.238	3.80	2.38–6.08	<.001
Proxy-Intensive Risk Score	0.956	0.205	2.60	1.74–3.89	<.001
Thin Credit File	0.624	0.187	1.87	1.30–2.69	.001
Human Individualized Review	−0.879	0.176	0.42	0.30–0.59	<.001
Data Minimization	−0.713	0.168	0.49	0.35–0.68	<.001
Applicant Appeal Access	−0.799	0.182	0.45	0.31–0.64	<.001
Explanation Requirement	−0.511	0.164	0.60	0.43–0.83	.002

Table 3 points to ambiguous criminal-history matching as the best positive predictor (odds ratio 3.80) of the aggressive adverse decision category a high risk label. Ambiguous evictions also tend to be associated with this label (with an odds ratio of 3.37), and a high risk score increases the odds of an aggressive adverse outcome by more than twice modeled risk. Thin credit files are also found to push the risk higher, though not as strongly.

The safeguard measures show the opposite trend. Human individualized review has the highest modeled protection with an odds ratio of 0.42. While other factors, Access of applicant to appeal has an OR of 0.45, Data minimization has an OR of 0.49, and Explanation requirements has an OR of 0.60. This leads to the conclusion that fairness can be improved not only by adjusting the algorithm, but by adjusting the treatment.

Graph 2: Simulated Adverse-Decision Risk Before and After Governance Controls



Graph 2 shows that many of the gains from structured human review occur in the very instances with the highest ambiguity of source. Clean-record instances drop from a modeled high-risk adverse-decision rate of 5.4% for unreviewed cases down to 2.1%. Thin-credit instances drop from 12.8% to 5.6%. And the gains finally grow larger for the eviction-record ambiguity, where risk drops from 18.9% to 7.4%.

Criminal-history ambiguity equates to 21.7% automated worst-case risk, reduced to 8.2% after adaptive review. This indicates a mechanism for adaptive oversight: here, where regular applications are predicated on reliable, verified information, a significant degree of automation is desirable. When court record ambiguity, identity mismatch uncertainty and score borderline are present, more human moderation is necessary.

Discussion / Critical Discussion

a) Interpretation of the Results

Overall, the key finding is that aggregate performance does not tell the whole story of how a given tenant-screening architecture makes its predictions. An AI-Only architecture yields a seemingly good accuracy of 91.2%, but the breakdown across subgroups shows an 8.8 point difference in false denials. In other words, a system can perform well by certain measures but misallocate its errors.

The Responsible-AI setting is operating differently because it does not learn that all historical information is equally accurate. Instead of translating questionable records into a risk penalty, it inquires if the record is verified, relevant, recent enough, and associated with a legitimate tenancy issue.

This distinction is relevant because administrative records are not neutral representations of underlying behavior. Records are driven by institutional processes, reporting practices, legal systems, data-quality concerns, and historical injustices. The work of Barocas and Selbst on disparate impact sheds light here: data-driven systems may simply be replicating existing social inequality regardless of whether they explicitly use protected-class attributes.

Consequently, the shrinkage from an 8.8-point false-denial gap to a 0.9-point false-denial gap must not be mistaken for evidence that Responsible AI has simply become fairer for all groups. Chouldechova and Kleinberg and others have shown that fairness metrics can conflict in mathematical space; a system that meets the requirements for one subgroup may not meet it for another. A variety of subgroup criteria must therefore be pursued.

b) Comparison With Previous Studies

The results echo longstanding difficulties with tenant-screening documented by Dunn and Grabchuk who recount the impact of background checks and the difficulty of tenant-screening data. The simulation raises this issue in automated scoring by illustrating how ambiguity can be translated into reproducible negative results in the absence of human context.

Reosti's study is also noteworthy in that it illustrates that landlord screening is still a form of discretionary decision making and not an unbiased technical task. If automation does eliminate some types of inconsistent judgement, then it relocates discretion into decisions

about ‘model specification; what to include in the model, how to weight the returns, which thresholds to set, and whether to take account of mitigation evidence.’

The simulation’s focus on subgroup false-denial rates adheres to a strong sense of equality of opportunity, as has been advanced by Jack Hardt and others. A housing applicant who is truly qualified by some accepted criterion should not be substantially more or less likely to be erroneously denied by a model using group membership as a predictor.

The results further justify Selbst and others’ criticism of purely technical fairness: “Housing, credit, and employment screening is not only a prediction problem. It is also a problem of applicants, landlords, screening firms, courts, reporting agencies, housing law, dispute mechanisms, and differential access to documentation. Any fairness intervention operating only within the classifier cannot account for these institutional relationships.

c) Why the Responsible-AI Model Performs Differently

The Responsible-AI architecture works on a different principle: uncertainty is treated as information in responsible-AI, and as guilt in the other architecture.

An AI-Only model would be likely to treat an eviction filing as a negative indicator even if the final disposition is unknown. A governed model would be asking the record Does this reliably establish the fact this model is trying to predict?

A thin credit file could just as likely mean a person has not engaged with the domestic credit markets as means of payment as a sign of an inability to satisfy rental obligations. Responsible schemes would avoid or not assume that absent conventional information automatically equates to high risk.

The three design features identified are responsible for the majority of the improvements.

First, minimization of data mitigates the use of substitute- or proxy-intensive variables.

Second, the structured human review provides context for uncertain records or records with weak model confidence.

Third, applicant contestability permits resolution of inaccurate or incomplete information prior to a final Housing Exclusion.

These stories show that the Responsible-AI system is willing to sacrifice a marginal amount of precision over the group when it can, in turn, greatly lower the number of false denials that discriminate.

Research Contribution / Implications

Theoretically, the study makes the contribution of differentiating algorithmic fairness, evidentiary fairness and procedural fairness.

The problem of algorithmic fairness deals with the statistical distribution of a predictive system over groups of applicants.

For evidentiary fairness, the question is: what has been entered into the system? To ask another way, does the information that is being presented in a manner that makes sense and that is correct, relevant, and comprehensive?

Procedural fairness: Did the applicant understand reasons, were the human considerations relevant, and did the applicant have a realistic opportunity to correct information?

The meaning of this for housing providers is significant. A screening system has to stand the test of both the vendor's inference of its prediction accuracy and the following:

- The set of variables used in the model
- The origin and reliability of those variables
- False-denial rates by subgroup
- Treatment of the missing data
- Refusing to include, or the treatment of eviction filings without clear dispositions
- record-matching procedures
- explanation quality
- human-override procedures
- Dispute and appeal workflows
- And monitoring once in operation

For tenant-screening organizations, Model Card-type documentation can be a good way to record intended use, model limitations, the data sets it was tested on, performance measures, and known failure modes.

For policymakers, this points to the fact that prohibiting explicit use of protected characteristics is not enough. Discrimination can occur through proxies, historical disparities or unequal distribution of errors, even if the model has never been told explicitly to discriminate. Barocas and Selbst provide the larger conceptual basis for this concern.

For researchers, the study suggests that the future of tenant-screening audits should include reporting of more than just overall accuracy and the area-under-the-curve statistics. It is

recommended that false denials, false approvals, calibration, threshold sensitivity, appeal correction, and data-quality errors for subgroups should be reportable in addition to usual predictive measures.

Human oversight is another consideration of the model presented here. Human review should never become an empty ritual. Human reviewers should be able to overrule the system and should have information available to understand the system's decision.

Conclusion

Automated tenant screening is also a governance challenge and provides an opportunity. While it's reassuring that algorithms may help bring consistency to certain aspects of a rental-application process, and be a more efficient use of human time, there are likely areas, such as identifying the most apt tenancies, that will be improved by a computer. However, they will also reinforce an existing, imperfect, perhaps even biased, record by further magnifying and legitimizing false information and be opaque to application in the form of gatekeeping.

The simulation built by this study illustrates the deficiencies of relying on aggregate accuracy as a measure of screening efficacy. The AI-Only model yielded the greatest total accuracy, while also had the largest modeled false-denial gap. Explainability improved transparency, but did not unilaterally eliminate disparate wrongs. Fairness-aware optimization reduced subgroup gap considerably, but the best overall fairness profile was attained by the combination of statistical controls and human review, data minimization, verified evidence, explanation, and appeal access.

Thus, the Responsible-AI architecture embodies an alternative approach to automation. There is no presumption that all screening will be done via human decision, nor the automated classifications will be determinative. Instead, the system optimizes the use of resources by efficiently processing routine cases and subjecting unclear or important cases to more extensive review.

The results further emphasize an important conceptual distinction between prediction and justification. A model may confidently predict negative consequences, but, nonetheless, housing exclusion must be justified with a defensible rationale derived from reliable, relevant, and properly understood evidence. Statistical confidence does not replace procedural fairness.

The figures in this fictional simulation are not an empirical estimate of discrimination in the housing markets. Many empirical studies of housing discrimination are based on synthetic data as well, and this paper's figures should be viewed as a demonstration of methods, not a true housing discrimination measure. An empirical test of this prediction remains to be done. However, the framework laid out here can be used as a basis to assess automated tenant screening as a sociotechnical system, one in which human threat and opportunity extend beyond integrated models and data to institutional incentives and institutional accountability.

Limitations

The main limitation is that the data used are simulated. There were no applicant features, no landlord decisions, no tenant-screening summaries, no protected-class information, and no housing results examined. The subgroup differences found cannot be applied to real markets. Both synthetic comparison groups were artificially generated for audit purposes and do not represent any actual racial, ethnic, disability, sex, family-status, or socioeconomic group.

A second limitation is the parsimonious set of criteria used to define tenant qualification. Actual landlords utilize various combinations of income, rental history, credit information, references, occupancy needs, and specific property considerations. Future research needs to identify legitimate criteria for tenancy vs. criteria to which predictive power (or legal propriety) is lacking.

Third, the simulation does not simulate all the fairness metrics. Equalized odds, predictive parity, demographic parity, calibration, individual fairness and counterfactual fairness can lead to different results.

Fourth, for the simulated data, the logistic-regression results are also simulated and should not be normalized as expression of causal effect.

Fifth, there is no legal advice. House and consumer-reporting duty can vary based on placement, property, program, and changing requirements of law and regulation.

Future Work

That future research examines automated tenant-screening systems using real data derived from legally and ethically obtained under privacy-concerned conditions.

Priority areas include:

- Auditing the false-denial rates across any truly defined set of comparison groups
- Evaluating the efficacy of automated matching of tenants to eviction records
- Separately the management of thin-credit applicants from that of applicants with adverse credit performance
- Assess the process for calculating income associated with the use of vouchers
- Frequency with which a human reviewer rejects the system-generated suggestion
- Assessing the effectiveness and availability of appeals available to applicants
- Study 1, testing whether explanation quality led to correction of decision
- Studying fairness across various rental-market environments.

Another avenue for future experimentation is to compare the effectiveness of different fairness interventions along various points of the machine-learning lifecycle: pre-processing, in-processing, adjusting decision thresholds, and post-decision auditing.

There should be more investigation into whether applicants understand the reason codes, and whether landlords rely too heavily on the automated recommendations when the screen explicitly states the recommendation is only a suggestion.

Likewise, because the system model may perform acceptably on subgroups it was designed to, it needs to be tested continuously. A model which fair at one point in its development can become and unfair as housing markets, applicant populations, court record systems, or credit behavior change.

Finally, research on tenant screening needs to consider fair-ness holistically in a housing ecology. Automated scoring engage with advertising, application fee practices, credit bureau reporting, eviction and landlord entry practices, rental supply, landlord laissez-faire, and rent applicant access to legal counsel. A technically better classifier cannot alone address inequality in other parts of this ecology.

Declarations

- **Conflict of Interest:** The hypothetical sample author declares no conflict of interest related to this methodological simulation. This statement must be replaced with the actual author's disclosure before publication.

- **Funding:** No external funding is attributed to this simulation-based manuscript. Any grant, institutional support, commercial sponsorship, or financial relationship associated with an empirical version of the study must be disclosed.
- **Ethics Statement:** The investigation used exclusively synthetic applicant records and did not involve human participants, real tenants, landlords, protected-class data, credit records, criminal-history records, eviction files, or personally identifiable information. Institutional human-subject approval was therefore not required for the simulation. Future empirical studies should obtain applicable institutional approval or exemption.
- **Data Availability:** All reported numerical outcomes were generated for methodological demonstration. No confidential, proprietary, commercial tenant-screening, consumer-reporting, or housing dataset was used.

References

1. Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104, 671–732. Retrieved from <https://doi.org/10.2139/ssrn.2477899>
2. Bhatia, S. (2020). Are tenant screening companies liable under the Fair Housing Act? *Fordham Law Review*, 88. retrieved from <https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=5689&context=flr>
3. 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>
4. Desmond, M., & Gershenson, C. (2017). Who gets evicted? Assessing individual, neighborhood, and network factors. *Social Science Research*, 62, 362–377. <https://doi.org/10.1016/j.ssresearch.2016.08.017>
5. Dunn, E., & Grabchuk, M. (2010). The social implications of case-by-case background checks: Modern residential tenant screening issues in Washington State. *Seattle Journal for Social Justice*, 9, 319–399. <https://digitalcommons.law.seattleu.edu/sjsj/vol9/iss1/13>
6. Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 329–338). <https://doi.org/10.1145/3287560.3287589>

7. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Proceedings of Advances in Neural Information Processing Systems*, 29, 3315–3323. <https://arxiv.org/abs/1610.02413>
8. Jagielski, M., Kearns, M., Mao, J., Oprea, A., Roth, A., & Sharifi-Malvajerdi, S., & Ullman, J. (2019). Fair fair learning under the differential privacy constraints. In *Proceedings of the 36th International Conference on Machine Learning*. <http://arxiv.org/abs/1812.02696>
9. Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. *Leibniz International Proceedings in Informatics*, 67, 43:1–43:23. <https://doi.org/10.4230/LIPIcs.ITCS.2017.43>
10. Kleinberg, J., & Mullainathan, S. (2019). Simplicity deepens inequality: How do we address biases, stereotypes, and interpretability. <https://arxiv.org/abs/1809.04578>
11. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the conference on Fairness, Accountability and Transparency* (pp. 220–229). [doi:10.1145/3287560.3287596](https://doi.org/10.1145/3287560.3287596)
12. Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). mitigating bias in algorithmic hiring: Evaluating claims and practices. in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. <https://arxiv.org/abs/1906.09208>
13. Reosti, A. (2020). “We go totally subjective”: Discretion, discrimination, and tenant screening in a landlord’s market. *Law & Social Inquiry*, 45(3), 618–657. <https://doi.org/10.1017/lsi.2019.81>
14. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59–68). <https://doi.org/10.1145/3287560.3287598>
15. U.S. Department of Housing and Urban Development. (2016). *Office of General Counsel Guidance on application of fair housing act standards to the use of criminal records by providers of housing and real estate-related transactions*. Retrieved from

<https://www.hkplawfirm.com/wp-content/uploads/sites/1401596/2019/10/HUD-April-2016-Memo-Criminal-Background-Checks.pdf>

16. U.S. Federal Trade Commission. (2018). *\$3 million FCRA settlement puts tenant background screening in the forefront*. Retrieved from <https://www.ftc.gov/business-guidance/blog/2018/10/3-million-fcra-settlement-puts-tenant-background-screening-forefront>
17. Woodworth, B., Gunasekar, S., Ohannessian, M. I., & Srebro, N. (2017). Learning non-discriminatory predictors. Retrieved from <https://arxiv.org/abs/1702.06081>
18. Boeing, G. (2020). Representation of online rental housing market and the digital reproduction of urban inequality. *Environment and Planning A: Economy and Space*. Retrieved from <https://arxiv.org/abs/1907.06118>