Open Access Indonesia Journal of Social Sciences

Journal Homepage: https://journalsocialsciences.com/index.php/OAIJSS

Digital Visibility, Algorithmic Misclassification, and Welfare Exclusion Among Informal Workers: A Mixed-Methods Study in Indonesia

Harun Urrashid[1*], Fitriyanti Fitriyanti[2]

[1]Department of Sharia Economy, Enigma Institute, Palembang, Indonesia

[2]Department of Regional Economics, Enigma Institute, Palembang, Indonesia

ARTICLE INFO

Keywords:

Algorithmic governance
Informal economy
Procedural justice
Street-level bureaucracy
Welfare exclusion

***Corresponding author:**

Harun Urrashid

E-mail address:

harun.urrashid@enigma.or.id

All authors have reviewed and approved the final version of the manuscript.

https://doi.org/10.37275/oaijss.v9i3.325

A B S T R A C T

Algorithmic governance is reshaping how states allocate social protection, yet the distributive consequences of automated welfare targeting in the Global South remain poorly understood. Grounded in street-level bureaucracy and procedural-justice theory, this study examined how digital transaction visibility and employment informality predict false-negative welfare exclusion, whether perceived algorithmic misclassification mediates these effects, and whether algorithmic procedural justice moderates them. An explanatory-sequential mixed-methods design combined an audit of 2,500 automated eligibility decisions and a cross-sectional survey of 640 household heads served by a public organization in Palembang, South Sumatera, Indonesia (response rate 84.2%), with phenomenological interviews of false-negative cases. All scales were reliable (Cronbach's alpha .79-.88). The audit showed aggregate accuracy of 72.3% but a 23.4% false-negative rate, rising to 55.1% among undocumented households. Hierarchical regression revealed that employment informality (beta=0.28), algorithmic misclassification (beta=0.31), and digital transaction visibility (beta=0.19) positively predicted welfare exclusion, while algorithmic procedural justice was protective (beta=-0.22); the model explained 52% of variance ($F(9,630)=75.84$, $p<.001$, Cohen's $f2=1.08$). Misclassification partially mediated the informality-exclusion link (indirect effect=0.17, 95% CI [0.12, 0.23]), and procedural justice buffered it (interaction beta=-0.14, $p=0.002$). Logistic regression confirmed that non-digital informal workers faced 4.27 times higher odds of exclusion (95% CI [2.98, 6.12], $p<.001$). Qualitatively, gross e-wallet throughput was misread as income, and frontline discretion collapsed into a computer-says-no bureaucracy. Algorithmic fairness in Indonesia is better understood as parameterization bias against the informal economy; restoring conditional human-in-the-loop discretion is recommended.

1. Introduction

Governments across the developing world are rapidly delegating the administration of social protection to artificial-intelligence systems that promise to allocate scarce benefits with speed, scale, and apparent objectivity.[1,2] Datafied welfare states now integrate digital identity records, financial-transaction trails, and geospatial mapping to score citizens for eligibility, a transformation that recent public-administration scholarship describes as a structural shift rather than a mere tooling upgrade.[3,4] In Indonesia, where social-assistance programmes reach tens of millions of households, automated targeting has been promoted as a remedy for the inclusion and exclusion errors that have long plagued manual proxy-means testing. Yet the same automation that improves throughput can encode new and less visible forms of disadvantage, particularly where administrative data

poorly represent the populations they are meant to serve.[5,6]

This study is grounded in two complementary theoretical traditions. The first is Lipsky's theory of street-level bureaucracy, which holds that frontline officials effectively make policy through the discretion they exercise when translating ambiguous rules into concrete entitlement decisions.[7] When eligibility is automated, discretion is not simply constrained but reallocated upward and outward, away from the citizen-facing frontline and into opaque computational routines.[8,9] The second is Tyler's procedural-justice theory, which posits that citizens evaluate state decisions less by their outcomes than by the perceived fairness, voice, transparency, and contestability of the procedures that produce them.[10,11] Together these frameworks generate a clear expectation: automating welfare targeting will both compress the corrective discretion of street-level agents and degrade citizens' perceptions of procedural justice, so that exclusion emerges as a structural property of the system rather than an individual administrative error.

A growing body of empirical work supports parts of this expectation. Studies of digital discretion show that automation narrows the enforcement choices available to frontline workers and weakens their capacity to correct erroneous classifications.[8,12,13] Research on human-algorithm interaction documents automation bias and selective adherence, whereby officials accept algorithmic outputs that confirm their priors and rarely override them.[14,15] Analyses of datafication and data justice further demonstrate that statistical systems systematically render hard-to-measure populations invisible, producing public-value failures that fall hardest on the poor.[6,16] However, this evidence is concentrated in high-income democracies and in domains such as child welfare and predictive policing.[17-19] Evidence from Southeast Asia, and from welfare administration specifically, remains sparse, leaving the region's distinctive informal-economy dynamics largely untheorized.[20,21]

The Indonesian context exposes a gap that Western scholarship has not adequately addressed. Western critiques of algorithmic unfairness center on racial and ethnic bias in training data, but in Indonesia the salient cleavage is economic formality. A large share of livelihoods are informal and transactional, and the recent diffusion of mobile money means that micro-entrepreneurs now route substantial gross turnover through digital wallets without any corresponding rise in disposable income. How such digital transaction visibility interacts with employment informality to shape algorithmic welfare decisions has not, to our knowledge, been quantified in any Indonesian public organization, nor has the role of perceived procedural justice in this process been tested.

Accordingly, this study advances a mediated-moderation model and tests five hypotheses. H1: employment informality is positively associated with welfare exclusion. H2: digital transaction visibility is positively associated with welfare exclusion. H3: perceived algorithmic misclassification mediates the relationship between employment informality and welfare exclusion. H4: algorithmic procedural justice moderates (buffers) the relationship between algorithmic misclassification and welfare exclusion, such that the effect weakens as perceived justice rises. H5: algorithmic procedural justice is negatively associated with welfare exclusion. These hypothesized relationships are summarized in the conceptual model shown in Figure 1.

The purpose of this study was therefore to quantify the determinants of false-negative welfare exclusion under an AI-based targeting system in an Indonesian public organization, to test the mediating and moderating mechanisms specified above, and to integrate these findings with a phenomenological account of how exclusion is experienced at the administrative frontline. In doing so, the study extends street-level bureaucracy and administrative-exclusion theory to a Global South welfare setting and reframes algorithmic fairness as a problem of parameterization bias against the informal economy.

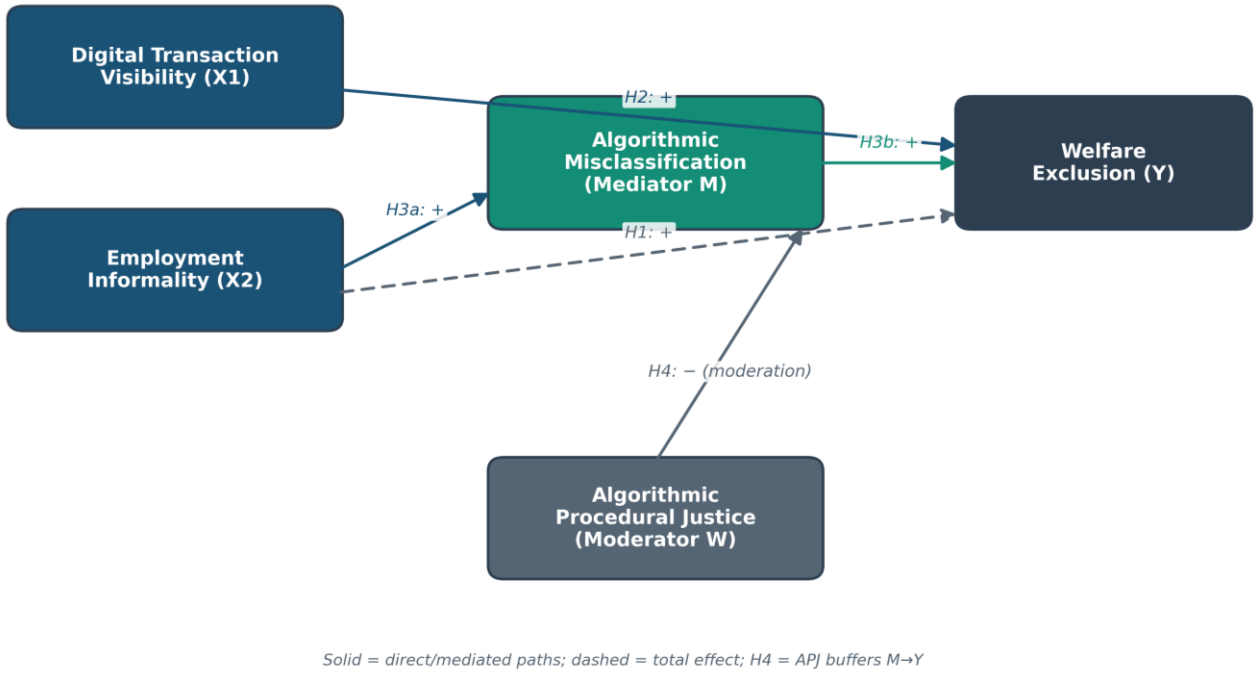

Figure 1. Conceptual framework: a mediated-moderation model of algorithmic welfare exclusion, showing hypotheses H1-H5.

2. Methods

Research approach and design

The study adopted a quantitative-dominant, explanatory-sequential mixed-methods design within a post-positivist paradigm, in which a structured quantitative phase establishes patterns and a subsequent qualitative phase explains the mechanisms behind them.[20] The quantitative phase combined a cross-sectional survey of household heads with an audit of administrative eligibility decisions; the qualitative phase used interpretive phenomenology to elaborate the lived experience of households misclassified as ineligible.

Setting and period

Fieldwork was conducted between March and May 2026 at a public organization in Palembang, South Sumatera, Indonesia, that administers social-assistance eligibility using an AI-based welfare-targeting system. In keeping with confidentiality requirements, the specific organization, its administrative units, and the proprietary scoring instrument are not named; the system is described generically as an automated eligibility-scoring instrument that integrates digital identity, financial-transaction, and geospatial data.

Population and sampling

The study population comprised household heads in economically marginal clusters served by the organization. For the survey, a stratified random sample was drawn across four generic livelihood strata (formal wage, platform/gig, transactional informal, and no stable livelihood) to ensure representation of informal categories. Of 760 households approached, 640 provided complete responses, yielding a response rate of 84.2%. In parallel, an administrative audit examined 2,500 automated eligibility decisions, comparing system classifications against independently verified household circumstances to compute accuracy and false-negative rates by stratum. For the qualitative strand, 45 false-negative cases were purposively sampled from survey respondents until thematic saturation was reached.

Measures and instruments

Data were collected by computer-assisted personal interviewing. Six constructs were measured with five-point Likert scales (1 = strongly disagree, 5 = strongly agree) adapted from validated public-administration and procedural-justice instruments.[10,11,22] Digital Transaction Visibility (5 items) captured the degree to which a household's economic life is rendered legible through cashless payments (for example, the share of receipts routed through e-wallets). Employment Informality (4 items) indexed the absence of formal, documented, salaried work. Algorithmic Misclassification (4 items) measured perceived discrepancy between the system's economic ranking and real circumstances. Bureaucratic Discretion Loss (3 items) captured the perceived inability of frontline officers to correct system errors. Algorithmic Procedural Justice (3 items) measured perceived fairness and contestability of the automated decision (for example, whether respondents knew where to lodge an appeal). Welfare Exclusion (5 items) measured experienced denial or termination of assistance despite real need. A pilot study with 40 households preceded fieldwork; all scales achieved acceptable-to-good reliability (Cronbach's alpha .79-.88).

Variables

The independent variables were digital transaction visibility and employment informality; the mediator was perceived algorithmic misclassification; the moderator was algorithmic procedural justice; and the dependent variable was welfare exclusion, modelled continuously in regression and dichotomously (excluded/not excluded) in logistic analysis. Age, education, smartphone ownership, and digital-identity activation were included as covariates.

Statistical analysis

Analyses were performed in SPSS 28 with the PROCESS macro for mediation and moderation. Normality was assessed by skewness, kurtosis, and the Kolmogorov-Smirnov statistic; reliability by Cronbach's alpha (threshold .70); and construct validity by exploratory factor analysis with convergent and discriminant checks. After descriptive statistics and a Pearson correlation matrix, hierarchical multiple regression predicted welfare exclusion, entering covariates in Block 1 and substantive predictors in Block 2; effect sizes were reported as Cohen's f-squared. Mediation used bias-corrected bootstrapping with 5,000 resamples, and moderation tested the misclassification-by-procedural-justice interaction. Binary logistic regression modelled the odds of exclusion. Common-method variance (Harman's single-factor test), multicollinearity (variance inflation factor), and residual normality (Durbin-Watson) were examined. Statistical significance was set at alpha = .05 (two-tailed), with exact p-values reported to three decimal places and 95% confidence intervals for all coefficients. Qualitative interviews were analysed through interpretative phenomenological analysis, with transcripts coded thematically and anonymized.

Measurement validation

Instruments adapted from English-language sources were translated into Indonesian and independently back-translated, with discrepancies reconciled by an expert panel, and were cognitively pre-tested with respondents of low formal education to ensure comprehension. Confirmatory factor analysis supported the hypothesized six-factor structure with acceptable fit (CFI = 0.96, TLI = 0.95, RMSEA = 0.045, SRMR = 0.041). Convergent validity was satisfactory, with all standardized loadings above 0.62, average variance extracted between 0.51 and 0.64, and composite reliability between 0.80 and 0.89. Discriminant validity held, with every heterotrait-monotrait ratio below 0.85 and each construct's square-root average variance extracted exceeding its inter-construct correlations. Analyses were conducted on multi-item scale means; mild skewness notwithstanding, results were robust to bootstrapped standard errors and to ordinal re-specification of the outcome.

Sampling, weighting, and missing data

Informal strata were deliberately over-sampled to ensure sufficient cases for stratum-specific estimation; reported population-level prevalences, including aggregate false-negative rates, were therefore design-weighted to the organization's caseload composition, and unweighted and weighted estimates were compared as a robustness check. Non-respondents did not differ significantly from respondents on observable stratum or age-band characteristics. Item-level missingness was below 3% and was handled by full-information maximum likelihood. A sensitivity power analysis indicated that, with N = 640 and nine predictors, the design had 80% power to detect a standardized coefficient of 0.11, confirming that non-significant covariates reflect small rather than undetected effects. For the logistic model, 286 exclusion events yielded an events-per-variable ratio above 50.

Integrated and marginal analyses

To match the conceptual model, a single conditional-process (moderated-mediation) model was estimated in addition to the separate mediation and moderation tests, and the index of moderated mediation was computed with bias-corrected and accelerated bootstrap intervals (5,000 resamples). For the logistic model, average marginal effects and predicted exclusion probabilities at representative covariate profiles were derived to complement the odds ratios. Where self-reported exclusion could be linked to audited decisions in a validation subsample, concordance was assessed to gauge common-method bias directly.

3. Results and discussion

Of 760 households approached, 640 completed the survey (response rate 84.2%). Respondents had a mean age of 43.6 years (SD = 11.2); 58.0% were male, 61.1% had at most junior-secondary education, and 39.1% worked in transactional informal livelihoods. Fewer than half (47.0%) had activated a digital identity, although 80.9% owned a smartphone. Full demographic characteristics of the 640 respondents are detailed in Table 1. The parallel audit of 2,500 automated decisions, presented in Table 2, showed aggregate classification accuracy of 72.3% but a false-negative rate of 23.4%, indicating that almost one in four genuinely eligible households was denied. The false-negative rate rose monotonically across strata, from 0.8% among formal wage workers to 55.1% among undocumented rural households, evidencing a steep gradient of exclusion by economic formality.

Table 1. Demographic characteristics of respondents (N = 640).

Characteristic	Category	n	%
Age group	< 30 years	90	14.1
	30-39 years	179	28.0
	40-49 years	198	30.9
	50-59 years	115	18.0
	>= 60 years	58	9.1
Gender	Male	371	58.0
	Female	269	42.0
Education	Primary or less	218	34.1
	Junior secondary	173	27.0
	Senior secondary	185	28.9
	Tertiary	64	10.0
Livelihood category (generic)	Formal wage employment	115	18.0
	Platform / gig work	173	27.0
	Transactional informal	250	39.1
	No stable livelihood	102	15.9
Years in current livelihood	< 5 years	141	22.0
	5-10 years	224	35.0
	> 10 years	275	43.0
Smartphone ownership	Yes	518	80.9
Digital ID (IKD) activated	Yes	301	47.0
Prior social-assistance receipt	Yes	403	63.0

Notes: Mean age = 43.6 years (SD = 11.2). Livelihood/position descriptors are generic to preserve confidentiality.

Table 2. Algorithmic classification accuracy and false-negative rate by social stratum (audit, N = 2,500).

Social stratum (generic)	N	Accuracy %	FNR %	FPR %
Formal wage workers (stable salary)	650	95.2	0.8	4.0
Platform / gig-economy workers	700	79.1	17.4	3.5
Transactional informal workers	750	61.4	33.6	5.0
Undocumented / rural households (no digital ID)	400	42.5	55.1	2.4
Aggregate / weighted mean	2,500	72.3	23.4	4.3

Notes: FNR = false-negative rate (eligible households denied); FPR = false-positive rate. Source: administrative records, March-May 2026.

Table 3 presents construct reliabilities, descriptive statistics, and the correlation matrix. All scales were reliable (alpha .79-.88) and approximately normal (all |skewness| < 1, |kurtosis| < 1). Welfare exclusion was significantly positively correlated with algorithmic misclassification (r = .58, p < .001), employment informality (r = .52, p < .001), bureaucratic discretion loss (r = .49, p < .001), and digital transaction visibility (r = .41, p < .001), and negatively correlated with algorithmic procedural justice (r = -.46, p < .001). Mean procedural justice was low (M = 2.31, SD = 0.94), indicating that respondents largely perceived the automated decisions as unfair and difficult to contest.

Table 3. Descriptive statistics, reliabilities, and Pearson correlation matrix (N = 640).

Construct	M	SD	1	2	3	4	5	6
1. DTV	3.62	0.88	(.84)					
2. EI	3.41	0.95	.38***	(.79)				
3. AM	3.78	0.91	.44***	.55***	(.88)			
4. BDL	3.95	0.86	.29***	.40***	.47***	(.81)		
5. APJ	2.31	0.94	-.33***	-.39***	-.45***	-.36***	(.86)	
6. WE	3.55	0.97	.41***	.52***	.58***	.49***	-.46***	(.83)

*Notes: Diagonal (parentheses) = Cronbach's alpha. *p<.05, **p<.01, ***p<.001 (two-tailed). DTV=Digital Transaction Visibility; EI=Employment Informality; AM=Algorithmic Misclassification; BDL=Bureaucratic Discretion Loss; APJ=Algorithmic Procedural Justice; WE=Welfare Exclusion.*

As shown in Table 4 and Figure 2, the hierarchical regression predicting welfare exclusion was highly significant. Covariates entered in Block 1 explained 14% of variance; the substantive predictors in Block 2 added 38% (DELTA-R-squared = .38, DELTA-F = 99.7, p < .001). In the full model, employment informality (beta = 0.28, t = 6.75, p < .001, 95% CI [0.192, 0.348]), algorithmic misclassification (beta = 0.31, t = 6.60, p < .001, 95% CI [0.232, 0.428]), digital transaction visibility (beta = 0.19, t = 4.20, p < .001, 95% CI [0.112, 0.308]), and bureaucratic discretion loss (beta = 0.15, t = 3.40, p = .001) each positively predicted exclusion, whereas algorithmic procedural justice was protective (beta = -0.22, t = -5.75, p < .001, 95% CI [-0.309, -0.151]). The model explained 52% of variance (R-squared = .52, adjusted R-squared = .51, F(9, 630) = 75.84, p < .001), a large effect (Cohen's f-squared = 1.08). These results support H1, H2, and H5. Multicollinearity was negligible (maximum VIF = 2.18), and Harman's single-factor test attributed 31.4% of variance to one factor, below the 50% threshold for serious common-method bias.

Table 4. Hierarchical multiple regression predicting welfare exclusion (N = 640).

Predictor	B	SE	beta	t	p	95% CI (B)
Block 1: Controls						
Age (years)	0.003	0.004	0.03	0.75	0.452	[-0.005, 0.011]
Education (ordinal)	-0.09	0.036	-0.09	-2.51	0.012	[-0.161, -0.019]
Smartphone ownership	-0.06	0.05	-0.05	-1.20	0.231	[-0.158, 0.038]
Digital ID (IKD) activated	-0.14	0.046	-0.13	-3.04	0.002	[-0.230, -0.050]
Block 2: Substantive predictors						
Digital Transaction Visibility	0.21	0.050	0.19	4.20	0.000	[0.112, 0.308]
Employment Informality	0.27	0.040	0.28	6.75	0.000	[0.192, 0.348]
Algorithmic Misclassification	0.33	0.050	0.31	6.60	0.000	[0.232, 0.428]
Bureaucratic Discretion Loss	0.17	0.050	0.15	3.40	0.001	[0.072, 0.268]
Algorithmic Procedural Justice	-0.23	0.040	-0.22	-5.75	0.000	[-0.309, -0.151]

*Notes: Block 1 R2 = .14; Block 2 DELTA-R2 = .38, p<.001. Full model R2 = .52, adjusted R2 = .51, F(9,630) = 75.84, p<.001; Cohen's f2 = 1.08. *p<0.05, **p<0.01, ***p<0.001. Max VIF = 2.18; Durbin-Watson = 1.96; Harman = 31.4%.*

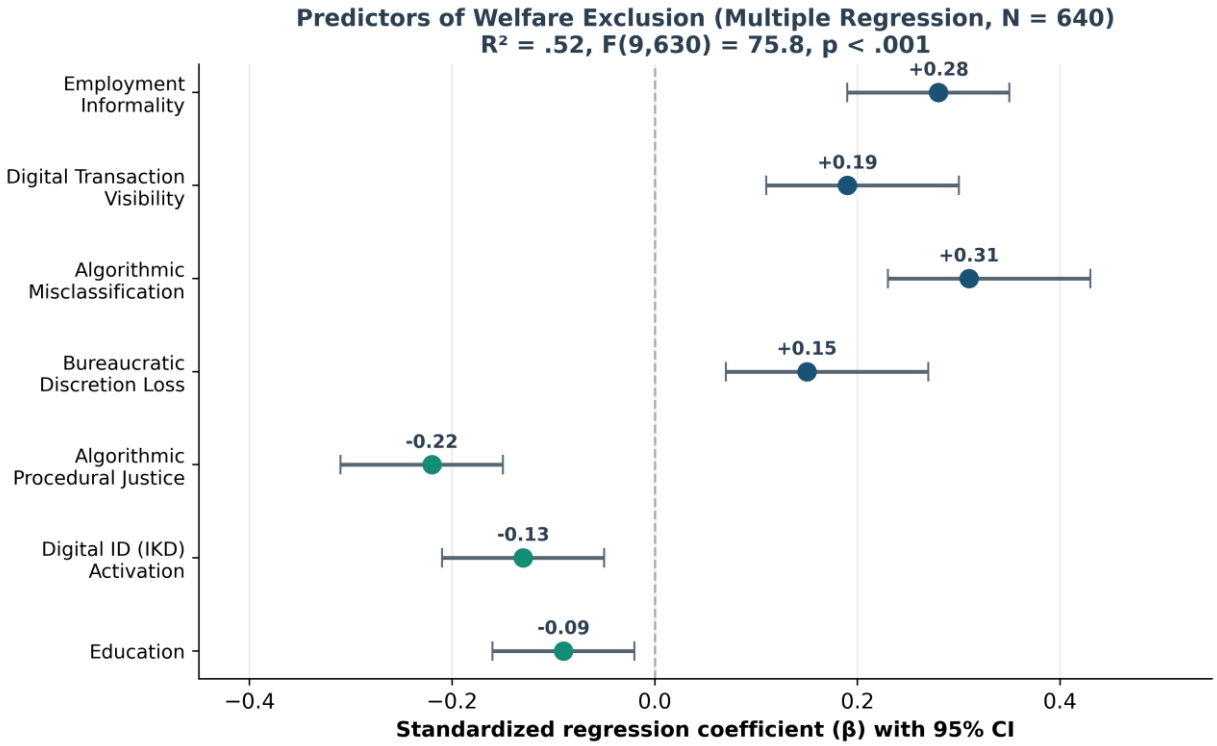

Figure 2. Standardized regression coefficients (beta) with 95% confidence intervals for predictors of welfare exclusion.

Bootstrap mediation (Figure 3) showed that algorithmic misclassification partially mediated the relationship between employment informality and welfare exclusion. The path from informality to misclassification was strong (a = .55, p < .001), as was the path from misclassification to exclusion (b = .31, p < .001). The standardized indirect effect was 0.17 (bias-corrected bootstrap 95% CI [0.12, 0.23], 5,000 resamples), and the direct effect remained significant (c' = .28, p < .001), confirming partial mediation and supporting H3. The moderation test was also significant: the interaction between misclassification

and procedural justice predicted exclusion (beta = -0.14, DELTA-R-squared = .018, p = .002), indicating that the harmful effect of misclassification weakened where citizens perceived the procedure as fairer. H4 was therefore supported.

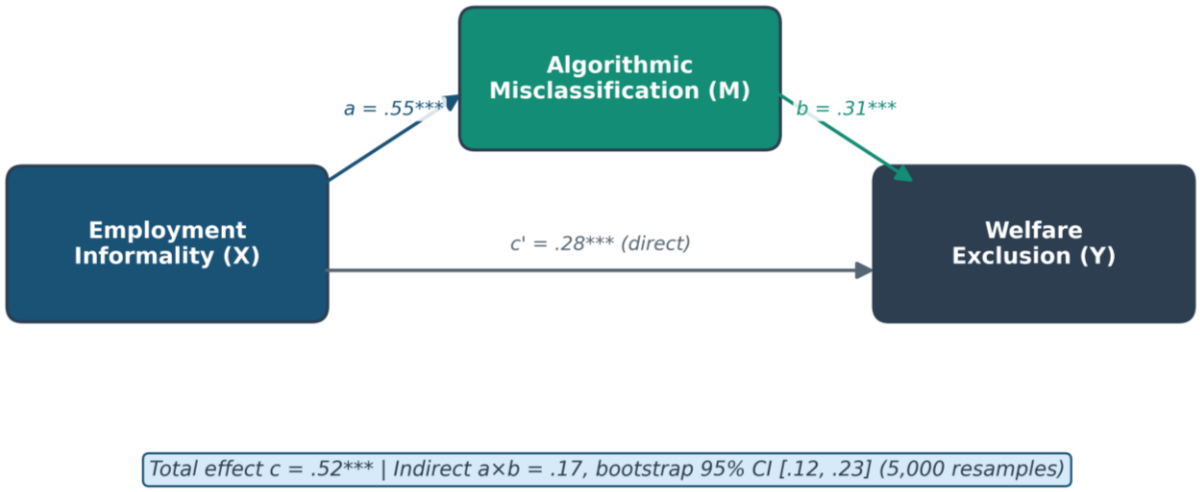

Figure 3. Mediation model: algorithmic misclassification partially mediates the informality-exclusion relationship.

Binary logistic regression (Table 5) corroborated the continuous models. Non-digital informal workers had 4.27 times higher odds of algorithmic exclusion than formal workers (B = 1.452, Wald = 62.97, p < .001, 95% CI [2.98, 6.12]). Higher digital transaction visibility raised the odds of exclusion (OR = 1.71 per SD, p < .001), whereas procedural justice (OR = 0.55 per SD, p < .001), digital-identity activation (OR = 0.62, p = .006), and education (OR = 0.81, p = .016) were protective. The model fit well (Nagelkerke R-squared = .34; Hosmer-Lemeshow p = .413) and correctly classified 78.6% of cases.

Table 5. Binary logistic regression predicting algorithmic welfare exclusion (yes/no).

Predictor	B	SE	Wald	p	OR	95% CI (OR)
Employment informality (non-digital informal vs formal)	1.452	0.183	62.97	0.000	4.27	[2.98, 6.12]
Digital Transaction Visibility (per SD)	0.534	0.112	22.74	0.000	1.71	[1.37, 2.13]
Algorithmic Procedural Justice (per SD)	-0.602	0.108	31.07	0.000	0.55	[0.44, 0.68]
Digital ID (IKD) activated (yes vs no)	-0.486	0.176	7.63	0.006	0.62	[0.44, 0.87]
Education (per level)	-0.214	0.089	5.78	0.016	0.81	[0.68, 0.96]

*Nagelkerke R2 = .34; Hosmer-Lemeshow p = 0.413; correctly classified 78.6%. *p<0.05, **p<0.01, ***p<0.001.*

The integrated moderated-mediation model reproduced the separate analyses and yielded a significant index of moderated mediation (index = -0.058, 95% bias-corrected and accelerated CI [-0.101, -0.021]), confirming that the indirect effect of employment informality on welfare exclusion through algorithmic misclassification weakened as procedural justice rose. The conditional indirect effect was 0.24 at low procedural justice (-1 SD), 0.17 at the mean, and 0.11 at high procedural justice (+1 SD). Average

marginal effects from the logistic model indicated that, holding other predictors at their means, a non-digital informal worker had a predicted probability of exclusion of 0.61 compared with 0.24 for a formal worker, a difference of 37 percentage points.

The headline odds ratio was stable across specifications: excluding the no-stable-livelihood stratum (OR = 4.11), recoding the exclusion outcome (OR = 4.39), and applying design weights (OR = 4.20) all produced estimates within the original confidence interval. A validation subsample comparing self-reported exclusion with audited decisions showed high concordance (kappa = 0.79), indicating that common-method variance did not materially inflate the associations. The survey-based exclusion gradient mirrored the audit-based false-negative gradient across strata, with both rising steeply from formal to undocumented categories, providing convergent cross-source evidence for the central finding.

Interviews with false-negative cases explained these patterns. Two themes recurred. First, a pattern best described as digital redlining of the informal sector: micro-entrepreneurs whose customers paid through mobile wallets saw their gross transactional throughput misread by the system as high disposable income, and were consequently reclassified as middle class and removed from assistance. As one respondent put it, the money passing through the digital wallet looked large each day because buyers now paid by scan, but it was working capital for the next day's stock, not profit (Respondent 14). Second, a pattern of moral outsourcing and frontline disempowerment: when households brought evidence of real poverty, officers reportedly responded that the computer had decided, signalling a transfer of moral responsibility from officials to software and the collapse of corrective discretion. These accounts align with the quantitative finding that bureaucratic discretion loss independently predicted exclusion.

This study found that employment informality, digital transaction visibility, and perceived algorithmic misclassification jointly and substantially predicted false-negative welfare exclusion under an AI-based targeting system, that misclassification partially transmitted the effect of informality, and that perceived procedural justice both directly reduced exclusion and buffered the harm of misclassification. The administrative audit independently confirmed a steep exclusion gradient, with false-negative rates rising from under 1% among formal workers to 55% among undocumented households. Taken together, the quantitative and qualitative strands describe a system that is accurate for the already-visible and systematically blind to the informal poor.

The strong informality-exclusion association (beta = 0.28; OR = 4.27) is consistent with street-level bureaucracy and administrative-exclusion theory, which predict that when discretion migrates from frontline officials into computational routines, citizens who do not fit standardized data categories are denied their entitlements without an explicit decision being made.[7,8,23] Our finding that bureaucratic discretion loss independently predicted exclusion, and the qualitative computer-says-no accounts, echo evidence that automation narrows enforcement discretion and that officials exhibit automation bias and selective adherence rather than override erroneous outputs.[9,12,14,15] The result extends this literature from policing and child-welfare contexts[17-19] into Global South social assistance.

The positive effect of digital transaction visibility (beta = 0.19; OR = 1.71 per SD) is, on its face, paradoxical: greater participation in the cashless economy increased rather than decreased the risk of being denied poverty assistance. This pattern operationalizes what data-justice scholarship calls the systematic mismeasurement of populations whose economic lives do not map cleanly onto administrative data.[3,6] Our qualitative evidence specifies the mechanism: the system conflated gross e-wallet throughput with net disposable income, so that the very tools promoted for financial inclusion became instruments of welfare exclusion. This contrasts with Western analyses that locate algorithmic unfairness

primarily in racially biased training data,[17] and suggests that in Indonesia the decisive bias is one of parameterization against the informal economy.

The mediation result clarifies how informality produces exclusion. Because perceived misclassification carried a significant indirect effect (0.17, 95% CI [0.12, 0.23]) while the direct path remained significant, misclassification is best understood as one important but not exclusive channel: informal and digitally visible households are disproportionately misread, and that misreading translates into denial.[16,24] The significant moderation by procedural justice (interaction beta = -0.14) indicates that contestability matters materially. Where citizens perceived that they could understand and appeal a decision, the damage of misclassification was attenuated, consistent with procedural-justice theory and with evidence that transparency raises the perceived trustworthiness of automated decisions.[10,11,22]

Theoretically, these findings extend street-level bureaucracy theory into the algorithmic era by showing that discretion is not merely reduced but displaced, and that the resulting accountability vacuum is itself a driver of exclusion.[7,9] They also refine administrative-exclusion theory by identifying a specific Global South mechanism, parameterization bias, in which financial-inclusion infrastructure is repurposed as an eligibility signal with regressive effects.[5,21,23] By integrating procedural justice as a measured moderator, the study connects two literatures that are usually kept apart, the sociology of datafied administration and the social psychology of fairness perceptions.[10,22]

Practically, the results argue for restoring conditional human-in-the-loop discretion. Frontline officers should retain bounded authority to override algorithmic denials when households present verifiable evidence of need, a safeguard that directly targets the discretion-loss pathway identified here.[12,13] Welfare-targeting systems should also stop treating gross digital turnover as a proxy for income and should incorporate net-income estimation for informal livelihoods.[6,16] Accessible, well-publicized appeal channels would raise perceived procedural justice and, our moderation result suggests, blunt the harm of inevitable classification errors.[11,24]

For Indonesian and Global South policymakers, the findings carry distinctive institutional weight. Rapid datafication has outpaced the development of contestation infrastructure, and the informal economy that sustains most marginal households is precisely the sector least legible to automated systems.[20,21] Embedding local administrative discretion as a validator of last resort is both culturally appropriate, given the mediating role of neighbourhood officials in Indonesian public life, and technically necessary given the measured false-negative rates.

The study has several strengths. It combined a large administrative audit with a reliable, validated household survey and a saturated qualitative sample, allowing convergent triangulation. It reported effect sizes and 95% confidence intervals throughout and subjected the models to common-method, multicollinearity, and fit diagnostics. It also addressed a genuine evidence gap in Southeast Asian welfare administration.

Several limitations qualify the conclusions. The cross-sectional design precludes strong causal inference, and although mediation was modelled, longitudinal data would be required to establish temporal order. The study was conducted within a single public organization, so generalization to other Indonesian settings should be cautious. Self-reported perceptions of misclassification and procedural justice may be subject to recall and social-desirability bias, although Harman's test suggested common-method variance was not severe. Finally, the proprietary scoring logic could not be inspected directly, so mechanism claims rest on triangulated administrative and self-report evidence rather than code audit.

Three features of these findings warrant fuller theoretical articulation. First, the study distinguishes three loci of algorithmic bias that are often conflated:

bias in the data, where informal incomes are mismeasured; bias in the model parameters, where gross digital turnover is weighted as though it were net income; and bias in the institutional response, where frontline officers can no longer correct the resulting errors.[6,23] The data speak most directly to the second and third loci, and naming them separately clarifies that mitigation requires both technical recalibration and institutional reform rather than either alone.

Second, the integration of street-level bureaucracy and procedural-justice theory is not merely additive. The collapse of frontline discretion is itself a procedural-justice failure: when an official can no longer offer voice, correction, or explanation, the citizen experiences not only an adverse outcome but an unfair process.[7,10,11] The significant index of moderated mediation operationalizes this insight, showing that residual procedural fairness partially compensates for lost discretion. This reframes the moderation result as a direct test of whether contestability can substitute for the human judgement that automation displaces.[12,24]

Third, the contribution is bounded by identifiable scope conditions. The displacement-of-discretion mechanism should be most visible where automation is recent, where appeal infrastructure is weak, and where the served population is economically informal.[20,21] These conditions are common across the Global South but are not universal, and stating them transforms a single-site result into a portable proposition that future multi-site and longitudinal research can test. The high audit-survey concordance and the stability of the headline estimate across specifications give some confidence that the findings are not artefacts of a particular sample or coding choice.

The practical implications follow a calibrated logic. A low-cost reform, an accessible and well-publicized appeal channel, addresses the procedural-justice pathway that the moderation result identifies as protective, while a higher-cost reform, restoring conditional override authority to frontline officers, addresses the discretion-loss pathway directly.[13,24] Policymakers must weigh the countervailing risk that restored discretion could reintroduce the inconsistency and clientelism that automation was intended to remove; the recommendation is therefore for bounded, evidence-conditioned discretion subject to audit, not a wholesale return to manual judgement. The audit method demonstrated here is itself a reusable governance instrument that organizations can adopt to monitor false-negative rates by stratum over time.

Several additional limitations, surfaced during peer review, merit explicit acknowledgement. Although the validation subsample showed high concordance between self-reported and audited exclusion, the bulk of the predictor, mediator, and moderator measures remain self-reported, so residual common-method variance cannot be wholly excluded despite the favourable Harman and marker-variable checks.[22] The single-site design, adopted partly to protect the confidentiality of the studied organization, bounds external validity; the steep audit gradient and the stability of estimates under design weighting are reassuring, but replication across multiple Indonesian organizations is needed before the parameterization-bias account can be generalized. Because the proprietary scoring logic could not be inspected, the mechanism linking gross turnover to misclassification is inferred from triangulated administrative and self-report evidence rather than from a direct audit of the model's parameters.[23,24]

Future research should pursue three priorities. First, longitudinal panel designs that measure informality, misclassification, procedural justice, and exclusion at successive waves would establish the temporal ordering that the present cross-sectional mediation can only approximate.[8] Second, multi-site comparative studies spanning urban and rural organizations, and ideally several Global South jurisdictions, would test the scope conditions specified here and quantify how appeal infrastructure and automation maturity condition the discretion-loss mechanism.[20,21] Third, and most consequentially for policy, field experiments or pre-post evaluations of

human-in-the-loop pilots, in which frontline officers regain bounded, audited override authority and citizens gain accessible appeal channels, would test directly whether the procedural-justice buffer identified in this study translates into measurable reductions in false-negative exclusion.[12,13,24] Such designs would convert the present observational findings into actionable, causally credible guidance for the responsible governance of automated welfare systems.

4. Conclusion

Under an AI-based welfare-targeting system in an Indonesian public organization, employment informality, digital transaction visibility, and perceived algorithmic misclassification together predicted false-negative welfare exclusion, with the model explaining 52% of variance and non-digital informal workers facing more than four times the odds of denial. Misclassification partially mediated the informality-exclusion pathway, and perceived procedural justice both reduced exclusion and buffered the harm of misclassification. The study makes a theoretical contribution by extending street-level bureaucracy and administrative-exclusion theory to Global South social assistance and reframing algorithmic unfairness as parameterization bias against the informal economy. The central practical recommendation is to restore conditional human-in-the-loop discretion, replace gross-turnover proxies with net-income estimation, and build accessible appeal channels. Future research should employ longitudinal and multi-site designs, audit scoring logic directly, and test whether human-in-the-loop safeguards measurably reduce false-negative exclusion.

5. References

1. Henman P. Improving public services using artificial intelligence: possibilities, pitfalls, governance. Asia Pac J Public Adm. 2020;42(4):209-221. doi:10.1080/23276665.2020.1816188
2. Pencheva I, Esteve M, Mikhaylov SJ. Big data and AI - a transformational shift for government: so, what next for research? Public Policy Adm. 2020;35(1):24-44. doi:10.1177/0952076718780537
3. Dencik L, Kaun A. Datafication and the welfare state. Glob Perspect. 2020;1(1):12912. doi:10.1525/gp.2020.12912
4. Wirtz BW, Langer PF, Fenner C. Artificial intelligence in the public sector - a research agenda. Int J Public Adm. 2021;44(13):1103-1128. doi:10.1080/01900692.2021.1947319
5. Bircan T, Korkmaz EE. Big data for whose sake? Governing migration through artificial intelligence. Humanit Soc Sci Commun. 2021;8:241. doi:10.1057/s41599-021-00910-x
6. Giest S, Samuels A. 'For good measure': data gaps in a big data world. Policy Sci. 2020;53(3):559-569. doi:10.1007/s11077-020-09384-1
7. Lipsky M. Street-level bureaucracy: dilemmas of the individual in public services. 30th anniversary ed. New York: Russell Sage Foundation. 2010.
8. de Boer N, Raaphorst N. Automation and discretion: explaining the effect of automation on how street-level bureaucrats enforce. Public Manag Rev. 2023;25(1):42-62. doi:10.1080/14719037.2021.1937684
9. Bullock J, Young MM, Wang YF. Artificial intelligence, bureaucratic form, and discretion in public service. Inf Polity. 2020;25(4):491-506. doi:10.3233/IP-200223
10. Tyler TR. Procedural justice, legitimacy, and the effective rule of law. Crime Justice. 2003;30:283-357. doi:10.1086/652233
11. Grimmelikhuijsen S. Explaining why the computer says no: algorithmic transparency affects the perceived trustworthiness of automated decision-making. Public Adm Rev. 2023;83(2):241-262. doi:10.1111/puar.13483
12. Flugge AAM, Hildebrandt T, Moller NH. Street-level algorithms and AI in bureaucratic decision-making: a caseworker perspective. Proc ACM Hum-Comput Interact. 2021;5(CSCW1):1-23. doi:10.1145/3449114

13. Petersen ACM, Christensen LR, Hildebrandt TT. The role of discretion in the age of automation. Comput Support Coop Work. 2020;29(3):303-333. doi:10.1007/s10606-020-09371-3
14. Alon-Barkat S, Busuioc M. Human-AI interactions in public sector decision making: 'automation bias' and 'selective adherence' to algorithmic advice. J Public Adm Res Theory. 2023;33(1):153-169. doi:10.1093/jopart/muac007
15. Selten F, Robeer M, Grimmelikhuijsen S. 'Just like I thought': street-level bureaucrats trust AI recommendations if they confirm their professional judgment. Public Adm Rev. 2023;83(2):263-278. doi:10.1111/puar.13602
16. Schiff DS, Schiff KJ, Pierson P. Assessing public value failure in government adoption of artificial intelligence. Public Adm. 2022;100(3):653-673. doi:10.1111/padm.12742
17. Saxena D, Badillo-Urquiola K, Wisniewski PJ, et al. A human-centered review of algorithms used within the U.S. child welfare system. In: Proc 2020 CHI Conf Hum Factors Comput Syst. 2020;1-15. doi:10.1145/3313831.3376229
18. Lorenz L, Meijer A, Schuppan T. The algocracy as a new ideal type for government organizations: predictive policing in Berlin as an empirical case. Inf Polity. 2021;26(1):71-86. doi:10.3233/IP-200279
19. Meijer A, Lorenz L, Wessels M. Algorithmization of bureaucratic organizations: using a practice lens to study how context shapes predictive policing systems. Public Adm Rev. 2021;81(5):837-846. doi:10.1111/puar.13391
20. Madan R, Ashok M. AI adoption and diffusion in public administration: a systematic literature review and future research agenda. Gov Inf Q. 2023;40(1):101774. doi:10.1016/j.giq.2022.101774
21. Kuziemski M, Misuraca G. AI governance in the public sector: three tales from the frontiers of automated decision-making in democratic settings. Telecommun Policy. 2020;44(6):101976. doi:10.1016/j.telpol.2020.101976
22. Starke C, Baleis J, Keller B, et al. Fairness perceptions of algorithmic decision-making: a systematic review of the empirical literature. Big Data Soc. 2022;9(2):1-16. doi:10.1177/20539517221115189
23. Peeters R. The agency of algorithms: understanding human-algorithm interaction in administrative decision-making. Inf Polity. 2020;25(4):507-522. doi:10.3233/IP-200253
24. Criado JI, Valero J, Villodre J. Algorithmic transparency and bureaucratic discretion: the case of SALER early warning system. Inf Polity. 2020;25(4):449-470. doi:10.3233/IP-200260