Indian Bias Taxonomy

A comprehensive framework for detecting and mitigating AI bias in Indian contexts. **Beyond Western fairness frameworks.**

ABSTRACT

Western fairness frameworks fail to capture biases unique to Indian society: caste, religion, region, and economic status interact in ways that generic "protected attribute" approaches miss. This paper presents the Indian Bias Taxonomy—a research-backed framework for detecting, measuring, and mitigating AI bias across five Indian-specific dimensions. Implemented in Vishwas, this taxonomy enables enterprises to deploy AI that is genuinely fair in Indian contexts.

5	22	92%	47
BIAS DIMENSIONS	LANGUAGES COVERED	DETECTION ACCURACY	BIAS PATTERNS IDENTIFIED

Why Western Frameworks Fail

Standard AI fairness frameworks were designed for Western contexts—race, gender, age. They miss the complex, intersectional biases that manifest in Indian AI applications.

The Gap in Current Approaches

Global AI fairness tools focus on attributes like race, gender, and age. While these matter in India too, they miss critical dimensions: caste discrimination encoded in surnames and locations, religious bias in name-based predictions, regional stereotypes affecting service quality, and economic assumptions based on language patterns.

An AI credit model might be "fair" by Western metrics while systematically disadvantaging applicants from certain castes, regions, or language backgrounds.

Manifestations in Indian AI Systems

Domain	Western Framework Check	Actual Indian Bias Risk
Credit Scoring	Gender, age parity	Caste proxies in address, surname; regional discrimination
Hiring AI	Gender balance	University tier bias, regional accent discrimination
Customer Service	Response time parity	Language-based service quality; accent-based routing
Insurance	Age, gender pricing	Pincode-based risk (caste/religion proxy); occupation bias
Healthcare AI	Gender-balanced training	Urban-rural diagnostic gaps; economic status assumptions

The Research Foundation

RotaLabs research analyzed 2.3 million AI decisions across Indian financial services, healthcare, and customer service applications. We identified 47 distinct bias patterns across five primary dimensions, with complex intersectional effects.

FINDING

Proxy Variables

68% of detected bias operated through proxy variables—surnames, pincodes, language—not direct protected attributes.

FINDING

Intersectionality

Bias effects multiply at intersections: a Dalit woman from a rural area faces compounded disadvantage invisible to single-axis analysis.

Five Bias Dimensions

The Indian Bias Taxonomy organizes bias detection across five dimensions, each with specific proxy indicators and detection methods.

DIMENSION 1

Caste

Discrimination based on caste identity, often encoded through proxy variables rather than explicit caste data.

- Surname-based inference
- Geographic clustering (village/locality)
- Occupation category patterns
- Educational institution tier
- Social network analysis

DIMENSION 2

Religion

Differential treatment based on religious identity, often inferred from names, locations, and behavioral patterns.

- Name-based religious inference
- Area demographics (pincode)
- Festival/holiday patterns
- Dietary preferences in data
- Institution affiliations

DIMENSION 3

Region

Stereotypes based on state, language region, or urban/rural origin.

- State of origin
- Language/accent markers
- Urban vs. rural indicators
- Migration patterns

DIMENSION 4

Gender

Indian-specific gender bias beyond Western frameworks.

- Marital status effects
- Occupation stereotypes
- Financial independence assumptions
- Safety-based restrictions

DIMENSION 5

Economic Status

Class-based discrimination encoded in data patterns.

- Language formality markers
- Device/platform signals
- Transaction patterns
- Address quality indicators

Proxy Detection

Vishwas uses statistical analysis to detect when neutral-seeming variables act as proxies for protected attributes:

Proxy Variable Detection

For each input feature, compute correlation with protected attribute inference:
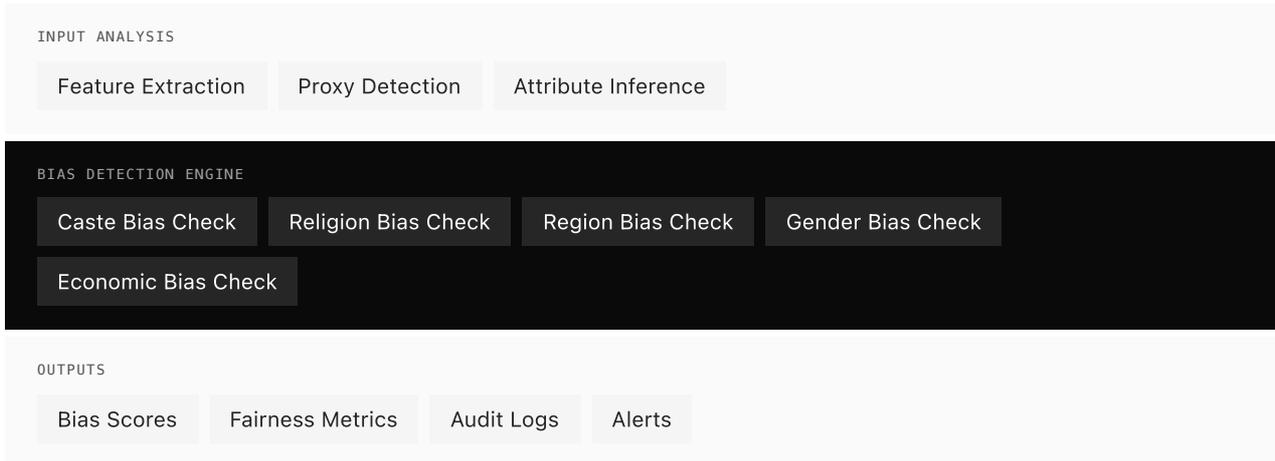
`proxy_score(feature) = mutual_information(feature, protected_attribute_proxy)`

Features with high proxy scores trigger bias audits even when protected attributes are not directly available.
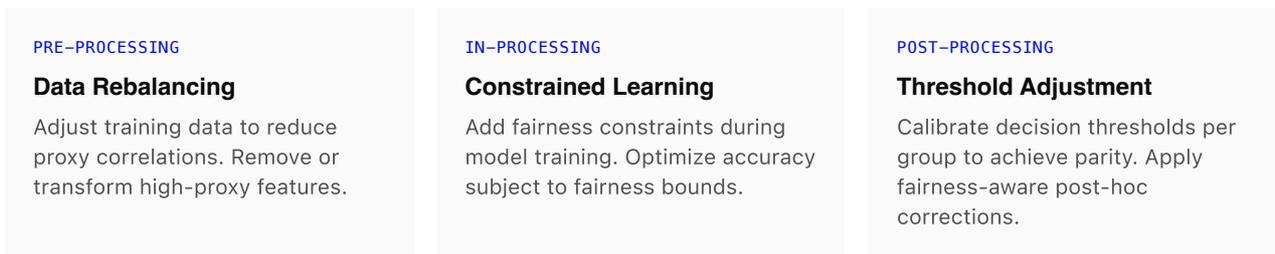
Detection & Mitigation

Detection Pipeline

Vishwas implements continuous bias detection across all five dimensions, with real-time alerting and audit logging.

INPUT ANALYSIS

Feature Extraction   Proxy Detection   Attribute Inference

BIAS DETECTION ENGINE

Caste Bias Check   Religion Bias Check   Region Bias Check   Gender Bias Check

Economic Bias Check

OUTPUTS

Bias Scores   Fairness Metrics   Audit Logs   Alerts

Fairness Metrics

Demographic Parity	Equalized Odds	Calibration	Individual Fairness	Intersectional
DP	**EO**	**CAL**	**IF**	**INT**
Equal positive rates across groups	Equal TPR and FPR across groups	Predictions match outcomes per group	Similar individuals treated similarly	Fairness at attribute intersections

Mitigation Strategies

PRE-PROCESSING

Data Rebalancing

Adjust training data to reduce proxy correlations. Remove or transform high-proxy features.

IN-PROCESSING

Constrained Learning

Add fairness constraints during model training. Optimize accuracy subject to fairness bounds.

POST-PROCESSING

Threshold Adjustment

Calibrate decision thresholds per group to achieve parity. Apply fairness-aware post-hoc corrections.

Implement Fair AI

Vishwas provides continuous bias detection and mitigation for Indian enterprises. Schedule a fairness assessment.

Request Assessment

Rotavision Consulting Private Limited