THE INDIA FORUM
A JOURNAL-MAGAZINE ON CONTEMPORARY ISSUES

March 10, 2026

Algorithms Don't Need Your Surname to Guess Your Caste. That's a Problem

By: Mohit Nirmender

A Dalit student graduating in the mid-2020s will encounter screening systems built on decades of unequal hiring, skewed promotion data, and biased public records. Unless institutions change how algorithmic systems are built, tested, and audited, inequality can only harden.

What's your surname? The question comes up everywhere. In a Gurgaon lift, over coffee in a Bangalore café, during a flat-hunting call in Delhi. It sounds polite, casual even. But in that moment, a stranger is running a background check to place you. Not through documents or references, but through your name alone: a caste certificate compressed into a single word.

Policy responses have focused on hiding this information. In 2021, a government-commissioned report recommended that UPSC interview panels should not know candidates' surnames The Bihar Public Service Commission adopted similar measures. The logic seemed sound: if surnames reveal caste and caste triggers bias, remove the surname. *Let merit speak for itself.*

When technology entered the hiring process, policymakers expected this logic to hold. Algorithms, after all, process only the data they're given. Remove the surname field, and the bias should disappear.

It didn't. The discrimination adapted. Algorithms trained on historical hiring data learnt to reconstruct caste information from variables that had nothing to do with names. The bias survived anonymisation because it was never really about the surname. It was about everything the surname predicted: address, school, father's occupation, gaps in employment. Remove one variable, and the algorithm finds ten others that correlate with it. In data science, this is called redundant encoding.

What follows isn't a story about surnames. It's an account of why institutional attempts to create neutrality fail when the data itself is a record of inequality.

How Algorithms Reconstruct Inequality

When researcher Dhiraj Singha asked ChatGPT to draft an academic cover letter in 2023, the AI quietly changed his surname from *Singha* to *Sharma*, a name far more common among dominant caste communities in Indian academia. This was not a random error. It reflected how the system had been trained.

ChatGPT learns by scanning large amounts of text and identifying which words tend to appear together. In academic documents, *Sharma* appears far more often in faculty lists, author bylines and institutional correspondence than *Singha* does. In that statistical environment, some surnames look like they belong in universities and others do not.

> By pushing caste out of view rather than confronting it, institutions create conditions in which bias survives in subtler forms and becomes easier to justify as neutral judgement.

More consequential discrimination occurs not in text generation but in decision systems. Applicant tracking systems used by Indian companies to screen résumés do not invent new names. They rank candidates by comparing profiles against historical hiring data.

This is where discrimination becomes structural.

In 2007, economist Sukhadeo Thorat and sociologist Paul Attewell sent identical job applications to over 500 companies across India, changing only the surnames (Thorat and Attewell 2007). Applicants with Dalit surnames received interview callbacks at roughly two thirds the rate of applicants with upper caste surnames. That study documented human bias. When such biased outcomes become the basis of hiring datasets, and those datasets are later used to train algorithms, discrimination stops being episodic. It becomes built into the system.

Consider what happens when a company attempts blind hiring. The surname field is removed from the application. The algorithm can no longer see *Sharma* or *Paswan*.

But it still sees residential address, postal code, undergraduate institution, father's occupation, employment gaps, extracurricular activities and even email domains. Each of these details is linked to caste history.

A candidate from a Kendriya Vidyalaya in a small town, whose father works as a low-level government servant, who has a two year gap after Class 12 and who lists no corporate internships is statistically more likely to come from a reserved category. A candidate from a private school in South Delhi, whose father's profession is listed as 'business', who studied abroad for a semester and who interned at a multinational is statistically more likely to be upper caste.

The algorithm does not need the surname. It reconstructs caste from what remains.

This pattern has been observed elsewhere. Research on algorithmic discrimination in Europe suggests that when one sensitive variable is removed, systems trained on historical data fall back on neutral looking signals such as neighbourhood, institutional background and past records, while continuing to reproduce group based disadvantage. Studies of surname bias in AI decision making show a similar logic. Pataranutaporn, Powdthavee, and Maes (2025) find that language models treat certain inherited family names as signals of intelligence, power and trustworthiness, and that these perceptions shape recommendations for hiring, leadership and loans even when qualifications are held constant.

> Concealing surnames does not neutralise discrimination. It reorganises it.

This is why the idea of neutral data is misleading. What algorithms learn are not objective measures of ability. They learn patterns produced by decades of institutional exclusion.

When an algorithm observes that candidates from elite colleges perform better in corporate roles than candidates from district colleges, it is not measuring talent alone. It is measuring who has historically been hired, promoted, mentored and retained. The result is a feedback loop. Upper caste candidates are hired more often, so they appear more frequently in datasets labelled "successful employees". The algorithm learns this correlation and then recommends more upper caste candidates.

What began as human discrimination hardens into algorithmic certainty. This logic extends beyond hiring. Government employment data show that Scheduled Caste professionals in India's IT sector earn about 25% less than colleagues with similar qualifications and job roles, and that employment probability gaps widened between 2011 and 2021 despite the sector's rapid growth. If such wage and employment data are used to train credit scoring systems, Dalit professionals are more likely to be classified as higher risk borrowers. Their lower incomes do not reflect individual creditworthiness. They reflect discriminatory hiring and promotion histories.

The algorithm does not see discrimination. It sees a pattern. AI does not invent caste bias. It treats historical inequality as information and reproduces it with the authority of numbers.

When Hiding Information Doesn't Work

The Indian state has tried to address surname-based discrimination by suppressing caste identifiers during interviews. Surnames are concealed, candidates are identified by roll numbers, and panels are instructed to avoid questions that reveal background, all in the name of protecting merit.

The effect has been different. Suppressing surnames has not produced clear evidence of altered selection outcomes. Instead, evaluation shifts away from explicit identifiers towards informal judgements that are harder to regulate. Interview panels increasingly rely on cues that appear neutral but are socially patterned, including conversational ease, linguistic style, and familiarity with elite institutional culture. These assessments carry no overt reference to caste, yet they reproduce its effects.

> The caste system has not disappeared. It has been rebuilt as infrastructure.

This shift matters because it changes how discrimination operates. Once caste moves out of formal criteria, it re-enters through qualities that cannot be easily audited or contested. Decisions are justified as matters of "fit", "confidence", or "communication skills", even when they follow long-standing social hierarchies. Discrimination becomes less visible, but more defensible.

Technology intensifies this process. When AI tools are introduced into interview screening, they inherit the same evaluative logic. Video-based systems assess facial expression, speech rhythm, and word choice. Research on facial recognition technologies shows that such systems absorb bias from their training data and perform worse on darker-skinned women (Buolamwini and Gebru 2018). Similar concerns apply to speech analysis. An algorithm trained primarily on corporate English spoken by upper-caste urban professionals will rank other speech patterns as weaker, not because they signal lower ability, but because they diverge from what the system has learnt to reward.

Concealing surnames, then, does not neutralise discrimination. It reorganises it. By pushing caste out of view rather than confronting it, institutions create conditions in which bias survives in subtler forms and becomes easier to justify as neutral judgement.

Structural Discrimination in Code

This is not only about individual prejudice. Even hiring managers who oppose caste discrimination can reproduce it when they rely on algorithmic tools trained on unequal historical data. The discrimination sits in the dataset, not in personal intent.

Evidence of this pattern is now visible. Journalists and researchers testing large language models with sentence-completion tasks have found that caste hierarchies reappear even when caste is not explicitly named. When prompted with phrases such as "The learned man is … ", models often suggest "Brahmin". Given "Do not touch the…" they return "Dalit" or related terms. When asked to generate lists of Indian doctors, professors, or experts, the outputs are dominated by upper-caste surnames. These are not isolated glitches. They reflect the social structure embedded in the data on which such systems are trained. News archives, institutional records, and professional listings disproportionately associate authority and expertise with upper-caste names, while Dalit names appear more often in contexts of violence, exclusion, or atrocity reporting. The algorithm does not invent this hierarchy. It learns how inequality is distributed in society and reproduces it as if it were a neutral map of reality.

> The difference now is that the judgement happens in milliseconds, behind interfaces that deny any social meaning at all.

When an algorithm penalises a Dalit candidate because historical data show lower employment or promotion outcomes for people with similar surnames or backgrounds, it treats the effects of past discrimination as predictions of individual ability. The candidate is judged not on their qualifications, but on what history has done to people like them.

The result is cumulative. A Dalit student graduating in the mid-2020s will encounter screening systems built on decades of unequal hiring, skewed promotion data, and biased public records. Rejection will appear data-driven and neutral, even as it reproduces exclusions that were never corrected. Unless institutions change how algorithmic systems are built, tested, and audited, inequality does not fade. It hardens.

What's in a Name?

Caste discrimination no longer requires explicit intent. It operates through systems that claim to optimise efficiency and objectivity. Decisions are made faster, at scale and without explanation. Responsibility dissolves into data.

And so the question remains. *What is your full name?* The answer still shapes what follows. The difference now is that the judgement happens in milliseconds, behind interfaces that deny any social meaning at all. The caste system has not disappeared. It has been rebuilt as infrastructure.

Mohit Nirmender is an independent researcher studying state institutions, public infrastructure, and social inequality in India.

References:

Buolamwini, Joy, and Timnit Gebru (2018). "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." *Proceedings of Machine Learning Research* 81: 1-15.

FRA (European Union Agency for Fundamental Rights) (2022). *Bias in Algorithms: Artificial Intelligence and Discrimination*. Luxembourg: Publications Office of the European Union.

MIT Technology Review (2025). "OpenAI Is Huge in India. Its Models Are Steeped in Caste Bias." October 1, 2025.

Pataranutaporn, Pat, Nattavudh Powdthavee, and Pattie Maes. 2025. "Algorithmic Inheritance: Surname Bias in AI Decisions Reinforces Intergenerational Inequality."

Chandran, Rina (2023) "India's Scaling Up of AI Could Reproduce Casteist Bias, Discrimination Against Women and Minorities." *Scroll.in* (14 September).

Sofi, Irfan Ahmad, Santosh Mehrotra, and Arun Kumar Bairwa (2024). "Myth of Meritocracy: Caste-Based Disparities in the IT Sector." *The Hindu* (25 December).

Thorat, Sukhadeo, and Paul Attewell (2007). "The Legacy of Social Exclusion: A Correspondence Study of Job Discrimination in India." *Economic and Political Weekly*. 42 (41): 4141-4145.