Re-imagining Algorithmic Fairness in India and Beyond

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, Vinodkumar Prabhakaran
(nithyasamba,erinarnesen,benhutch,tulsee,vinodkpg)@google.com
Google Research
Mountain View, CA

ABSTRACT
Conventional algorithmic fairness is West-centric, as seen in its subgroups, values, and methods. In this paper, we de-center algorithmic fairness and analyse AI power in India. Based on 36 qualitative interviews and a discourse analysis of algorithmic deployments in India, we find that several assumptions of algorithmic fairness are challenged. We find that in India, data is not always reliable due to socio-economic factors, ML makers appear to follow double standards, and AI evokes unquestioning aspiration. We contend that localising model fairness alone can be window dressing in India, where the distance between models and oppressed communities is large. Instead, we re-imagine algorithmic fairness in India and provide a roadmap to re-contextualise data and models, empower oppressed communities, and enable Fair-ML ecosystems.

CCS CONCEPTS
• Human-centered computing → Empirical studies in HCI.

KEYWORDS
India, algorithmic fairness, caste, gender, religion, ability, class, feminism, decoloniality, anti-caste politics, critical algorithmic studies

1 INTRODUCTION
Despite the exponential growth of fairness in Machine Learning (AI) research, it remains centred on Western concerns and histories—the structural injustices (e.g., race and gender), the data (e.g., ImageNet), the measurement scales (e.g., Fitzpatrick scale), the legal tenets (e.g., equal opportunity), and the enlightenment values. Conventional western AI fairness is becoming a universal ethical framework for AI; consider the AI strategies from India [1], Tunisia [205], Mexico [129], and Uruguay [13] that espouse fairness and transparency, instead of AI fairness failures and stakeholder coordination have resulted in bans and moratoria in the US. Several factors led to this outcome:
• Decades of scientific empiricism on proxies and scales that correspond to subgroups in the West [73].
• Public datasets, APIs, and freedom of information acts are available to researchers to analyse model outcomes [19, 113].
• AI research/industry is fairly responsive to bias reports from users and civil society [16, 46].
• The existence of government representatives glued into technology policy, shaping AI regulation and accountability [213].
• An active media systematically scrutinises and reports on downstream impacts of AI systems [113].

We argue that the above assumptions may not hold in much else of the world. While algorithmic fairness keeps AI within ethical and legal boundaries in the West, there is a real danger that naive generalisation of fairness will fail to keep AI deployments in check in the non-West. Scholars have pointed to how neoliberal AI follows the technical architecture of classic colonialism through data extraction, impairing indigenous innovation, and shipping manufactured goods back to the data subjects—among communities already prone to exploitation, under-development, and inequality from centuries of imperialism [39, 118, 137, 169, 211]. Without engagement with the conditions, values, politics, and histories of the non-West, AI fairness can be a tokenism, at best—pernicious, at worst—for communities. If algorithmic fairness is to serve as the ethical compass of AI, it is imperative that the field recognise its own defaults, biases, and blindspots to avoid exacerbating historical harms that it purports to mitigate. We must take pains not to develop a general theory of algorithmic fairness based on the study of Western populations. Could fairness, then, have structurally different meanings in the non-West? Could fairness frameworks that rely on Western infrastructures be counterproductive elsewhere?

How do social, economic, and infrastructural factors influence Fair-ML?

In this paper, we study algorithmic power in contemporary India, and holistically re-imagine algorithmic fairness in India. Home to 1.38 billion people, India is a pluralistic nation of multiple languages, religions, cultural systems, and ethnicities. India is the site of a vibrant AI workforce. Hype and promise is palpable around AI—envisioned as a force-multiplier of socio-economic benefit for a large, under-privileged population [1]. AI deployments are prolific, including in predictive policing [32] and facial recognition [61]. Despite the momentum on high-stakes AI systems, currently there is a lack of substantial policy or research on advancing algorithmic fairness for such a large population interfacing with AI.

We report findings from 36 interviews with researchers and activists working in the grassroots with marginalised Indian communities, and from observations of current AI deployments in India. We use feminist, decolonial, and anti-caste lenses to analyze our data. We contend that India is on a unique path to AI, characterised
by pluralism, socio-economic development, technocratic nation-building, and uneven AI capital—which requires us to confront many assumptions made in algorithmic fairness. Our findings point to three factors that need attention in Fair-ML in India:

**Data and model distortions**: Infrastructures and social contracts in India challenge the assumption that datasets are faithful representations of people and phenomena. Models are over-fitted for digitally-rich profiles—typically middle-class men—further excluding the 50% without Internet access. Sub-groups like caste (endogamous, ranked social identities, assigned at birth [17, 191]),

1 gender, and religion require different fairness implementations; but AI systems in India are under-analyzed for biases, mirroring the limited mainstream public discourse on oppression. Indic social justice, like reservations, presents new fairness evaluations.2

**Double standards and distance by ML makers**: Indian users are perceived as ‘bottom billion’ data subjects, petri dishes for intrusive models, and given poor recourse—thus, effectively limiting their agency. While Indians are part of the AI workforce, a majority work in services, and engineers do not entirely represent marginalities, limiting re-mediation of distances.

**Unquestioning AI aspiration**: The AI imaginary is aspirational in the Indian state, media, and legislation. AI is readily adopted in high-stakes domains, often too early. Lack of an ecosystem of tools, policies, and stakeholders like journalists, researchers, and activists to interrogate high-stakes AI inhibits meaningful fairness in India.

In summary, we find that conventional Fair-ML may be inappropriate, insufficient, or even inimical in India if it does not engage with the local structures. In a societal context where the distance between models and dis-empowered communities is large—via technical distance, social distance, ethical distance, temporal distance, and physical distance—a myopic focus on localising fair model outputs alone can backfire. We call upon fairness researchers working in India to engage with end-to-end factors impacting algorithms, like datasets and models, knowledge systems, the nation-state and justice systems, AI capital, and most importantly, the oppressed communities to whom we have ethical responsibilities. We present a holistic framework to operationalise algorithmic fairness in India, calling for: re-contextualising data and model fairness; empowering oppressed communities by participatory action; and enabling an ecosystem for meaningful fairness.

Our paper contributes by bringing to light how algorithmic power works in India, through a bottom-up analysis. Second, we present a holistic research agenda as a starting point to operationalise Fair-ML in India. The concerns we raise may certainly be true of other countries. The broader goal should be to develop global approaches to Fair-ML that reflect the needs of various contexts, while acknowledging that some principles are specific to context.

2 BACKGROUND

Recent years have seen the emergence of a rich body of literature on fairness and accountability in machine learning e.g., [29, 133]. However, most of this research is framed in the Western context, by researchers situated in Western institutions, for mitigating social injustices prevalent in the West, using data and ontologies from the West, and implicitly imparting Western values, e.g., in the premier FAccT conference, of the 138 papers published in 2019 and 2020, only a handful of papers even mention non-West countries, and only one of them—Marda’s paper on New Delhi’s predictive policing system[127]—substantially engages with a non-Western context.

## 2.1 Western Orientation in Fair-ML

2.1.1 **Axes of discrimination**. The majority of fairness research looks at racial [46, 57, 121, 125, 183] and gender biases [42, 46, 197, 219] in models—two dimensions that dominate the American public discourse. However these categories are culturally and historically situated [85]. Even the categorisation of proxies in fairness analyses have Western biases and origins; e.g., Fitzpatrick skin type is often used by researchers as a phenotype [46], but was originally developed to categorise UV sensitivity [73]. While other axes of discrimination and injustices such as disability status [91], age [59], and sexual orientation [76] have gotten some attention, biases relevant to other geographies and cultures are not explored (e.g., Adivasis (indigenous tribes of South Asia) and Dalits). As [142] points out, tackling these issues require a deep understanding of the social structures and power dynamics therein, which points to a wide gap in literature.

2.1.2 **Legal framing**. Since early inquiries into algorithmic fairness largely dealt with US law enforcement (predictive policing and recidivism risk assessment) as well as state regulations (e.g., in housing, loans, and education), the research framings often rely implicitly on US laws such as the Civil Rights Acts and Fair Housing Act, as well as on US legal concepts of discrimination. Indeed, researchers since the late 1960s have tried to translate US anti-discrimination law into statistical metrics [90]. The community also often repurposes terminology from US legal domains, such as “disparate impact”, “disparate treatment”, and “equal opportunity”, or use them as points of triangulation, in order to compare technical properties of fairness through analogy with legal concepts [81, 82].

2.1.3 **Philosophical roots**. Connections have been made between algorithmic fairness and Western concepts such as egalitarianism [38], consequentialism [142, 167], deontic justice [38, 142], and Rawls’ distributive justice [99, 142]. Indeed, notions of algorithmic fairness seem to fit within a broad arc of enlightenment and post-enlightenment thinking, including in actuarial risk assessment [147]. Dr. B. R. Ambedkar’s ((1891–1956), fondly called Babasaheb, the leader and dominant ideological source of today’s Dalit politics) anti-caste movement was rooted in social justice, distinct from Rawls’s distributive justice [166] (also see Sen’s critique of Rawl’s idea of original position and inadequacies of impartiality-driven justice and fairness [186]). Fairness’ status as the de facto moral standard of choice and signifier of justice, is itself a sign of cultural situatedness. Other moral foundations [80] of cultural importance may often be overlooked by the West, including purity/sanctity. Traditional societies often value restorative justice, which emphasises repairing harms [44], rather than fairness, e.g., contemporary Australian Aboriginal leaders emphasise reconciliation rather than

1 According to Shanmugavelan, “caste is an inherited identity that can determine all aspects of one’s life opportunities, including personal rights, choices, freedom, dignity, access to capital and effective political participation in caste-affected societies” [191]. Dalits (‘broken men’ in Maratha) are the most inferiorised category of the people, are within the social hierarchy, but excluded in caste categories [17, 35, 184, 191].

2 We use the term ‘Indic’ to refer to native Indian concepts.
fairness in their political goals [94]. Furthermore, cultural relationships such as power distance, and temporal orientation, are known to mediate the importance placed on fairness [122].

### 2.2 Fairness perceptions across cultures

Social psychologists have argued that justice and fairness require a lens that go beyond the Euro-American cultural confines [120]. While the concern for justice has a long history in the West (e.g., Aristotle, Rawls) and the East (e.g., Confucius, Chanakya), they show that the majority of empirical work on social justice has been situated in the US and Western Europe, grounding the understanding of justice in the Western cultural context. Summarising decades worth of research, [120] says that the more collectivist and hierarchal societies in the East differs from the more individualistic and egalitarian cultures of the West in how different forms of justice—distributive, procedural, and retributive—are conceptualised and achieved. For instance, [40] compared the acquisition of fairness behaviour in seven different societies: Canada, India, Mexico, Peru, Senegal, Uganda, and the USA, and found that while children from all cultures developed aversion towards disadvantageous inequity (avoid receiving less than a peer), advantageous inequity aversion (avoid receiving more than a peer) was more prevalent in the West. Similarly, a study of children in three different cultures found that notions of distributive justice are not universal: “children from a partially hunter-gatherer, egalitarian African culture distributed the spoils more equally than did the other two cultures, with merit playing only a limited role” [185]. See above point on Dr. B. R. Ambedkar’s centring on priority-based social justice for caste inequalities. The above works point to the dangers in defining fairness of algorithmic systems based solely on a Western lens.

### 2.3 Algorithmic fairness in the non-West

The call for a global lens in AI accountability is not new [84, 155], but the ethical principles in AI are often interpreted, prioritised, contextualised, and implemented differently across the globe [98]. Recently, the IEEE Standards Association highlighted the monopoly of Western ethical traditions in AI ethics, and inquired how incorporating Buddhist, Ubuntu, and Shinto-inspired ethical traditions might change the processes of responsible AI [93]. Researchers have also challenged the normalisation of Western implicit beliefs, biases, and issues in specific geographic contexts; e.g., India, Brazil and Nigeria [180], and China and Korea [192]. Representational gaps in data is documented as one of the major challenges in achieving responsible AI from a global perspective [21, 190]. For instance, [190] highlights the glaring gaps in geo-diversity of open datasets such as ImageNet and Open Images that drive much of the computer vision research. [12] shows that NLP models disproportionately fail to even detect names of people from non-Western backgrounds.

### 2.4 Accountability for unfairness

Discussion of accountability is critical to any discussions of fairness, i.e., how do we hold deployers of systems accountable for unfair outcomes? Is it fair to deploy a system that lacks in accountability? Accountability is fundamentally about answerability for actions [114], and central to these are three phases by which an actor is made answerable to a forum: information-sharing, deliberation and discussion, and the imposition of consequences [214]. Since outcomes of ML deployments can be difficult to predict, proposals for accountability include participatory design [109] and participatory problem formulation [128], sharing the responsibility for designing solutions with the community. Nissenbaum distinguishes four barriers to responsibility in computer systems: (1) the problem of many hands, (2) bugs, (3) blaming the computer, and (4) ownership without liability [146]. These barriers become more complicated when technology spans cultures: more, and more remote, hands are involved; intended behaviours may not be defined; computer-blaming may meet computer-worship head on (see Section 4); and questions of ownership and liability become more complicated.

Specific to the Indian context, scholars and activists have outlined opportunities for AI in India [104], proposed policy deliberation frameworks that take into account the unique policy landscape of India [126], and questioned the intrusive data collection practices through Aadhaar (biometric-based unique identity for Indians) in India [158, 159]. Researchers have documented societal biases in predictive policing in New Delhi [127], caste [202] and ethnicity [15] biases in job applications, call-center job callbacks [27], caste-based wage-gaps [124], caste discrimination in agricultural loans decisions [116], and even in online matrimonial ads [157].

### 3 Method

Our research results come from a critical synthesis of expert interviews and discourse analysis. Our methods were chosen in order to provide an expansive account of who is building ML for whom, what the on-the-ground experiences are, what the various processes of unfairness and exclusions are, and how they relate to social justice.

We conducted qualitative interviews with 36 expert researchers, activists, and lawyers working closely with marginalised Indian communities at the grassroots. Expert interviews are a qualitative research technique used in exploratory phases, providing practical insider knowledge and surrogacy for a broader community [41]. Importantly, experts helped us gain access to a nascent and difficult topic, considering the early algorithmic deployments in the public sector in India. Our respondents were chosen from a wide range of areas to create a holistic analysis of algorithmic power in India. Respondents came from Computer Science (11), Activism (9), Law and Public Policy (6), Science and Technology Studies (5), Development Economics (2), Sociology (2), and Journalism (1). All respondents had 10-30 years of experience working with marginalised communities or on social justice. Specific expertise areas included caste, gender, labour, disability, surveillance, privacy, health, constitutional rights, and financial inclusion. 24 respondents were based in India, 2 in Europe, 1 in Southeast Asia, the rest in the USA; 25 of them self-identified as male, 10 as female, and 1 as non-binary.

In conjunction with qualitative interviews, we conducted an analysis of various algorithmic deployments and emerging policies in India, starting from Aadhaar (2009). We identified and analysed various Indian news publications (e.g., TheWire.in, Times of India), policy documents (e.g., NITI Aayog, Srikrishna Bill), and community media (e.g., Roundtable India, Feminism in India), and prior research. Due to secondary sources, our citations are on the higher side.

**Recruitment and moderation** We recruited respondents via a combination of reaching out directly and personal contacts, using purposeful sampling [151]—i.e., identifying and selecting experts with relevant experience—iterative until saturation. We conducted all interviews in English (preferred language of participants). The
Semi-structured interviews focused on 1) unfairness through discrimination in India; 2) technology production and consumption; 3) the historical and present role of fairness and ethics in India; 4) biases, stereotypes and proxies; 5) data; 6) laws and policy relating to fairness; and 7) canonical applications of fairness, evaluated in the Indian context. Respondents were compensated for the study (giftcards of 100 USD, 85 EUR, and 2000 INR), based on purchasing power parity and non-coercion. Employer restrictions prevented us from compensating government employees. Interviews lasted an hour each, and were conducted using video conferencing and captured via field notes and video recordings.

**Analysis and coding** Transcripts were coded and analyzed for patterns using an inductive approach [201]. From a careful reading of the transcripts, we developed categories and clustered excerpts, conveying key themes from the data. Two team members created a code book based on the themes, with seven top-level categories (sub-group discrimination, data and models, law and policy, ML biases and harms, AI applications, ML makers, and solutions) and several sub-categories (e.g., caste, missing data, proxies, consent, algorithmic literacy, and so on). The themes that we describe in Section 4 were then developed and applied iteratively to the codes.

Our data is analysed using feminist, decolonial, and anti-caste lenses. A South Asian feminist stance allows us to examine oppressed communities as encountering and subverting forces of power, while locating in contextual specifics of family, caste, class, and religion. South Asian feminism is a critique of the white, Western feminism that saw non-western women as powerless victims that needed rescuing [49, 139]. Following Dr. B. R. Ambedkar’s insight on how caste hierarchies and patriarchies are linked in India [48], we echo that no social justice commitment in India can take place without examining caste, gender, and religion. South Asian feminism is a critique of the white, Western feminism that saw non-western women as powerless victims that needed rescuing [49, 139].

**Research ethics** We took great care to create a research ethics protocol to protect respondent privacy and safety, especially due to the sensitive nature of our inquiry. During recruitment, participants were informed of the purpose of the study, the question categories, and researcher affiliations. Participants signed informed consent acknowledging their awareness of the study purpose and researcher affiliation prior to the interview. At the beginning of each interview, the moderator additionally obtained verbal consent. We stored all data in a private Google Drive folder, with access limited to our team. To protect participant identity, we deleted all personally identifiable information in research files. We redact identifiable details when quoting participants. Every respondent was given the choice of default anonymity or being included in Acknowledgements. All co-authors of this paper work at the intersection of under-served communities and technology, with backgrounds in HCI, critical algorithmic studies, and ML fairness. The first author constructed the research approach and has had grassroots commitments with marginalised Indian communities for nearly 15 years. The first and second author moderated interviews. All authors were involved in the framing, analysis, and synthesis. Three of us are Indian and two of us are White. All of us come from privileged positions of class and/or caste. We acknowledge the above are our interpretations of research ethics, which may not be universal.

4 FINDINGS

We now present three themes (see Figure 1) that we found to contrast views in conventional algorithmic fairness.

4.1 Data and model distortions

Datasets are often seen as reliable representations of populations. Biases in models are frequently attributed to biased datasets, presupposing the possibility of achieving fairness by “fixing” the data [89]. However, social contracts, informal infrastructures, and population scale in India lead us to question the reliability of datasets.

4.1.1 Data considerations.

**Missing data and humans** Our respondents discussed how data points were often missing because of social infrastructures and systemic disparities in India. Entire communities may be missing or misrepresented in datasets, exacerbated by digital divides, leading to wrong conclusions [30, 52, 53, 119] and residual unfairness [103]. Half the Indian population lacks access to the Internet—the excluded half is primarily women, rural communities, and Adivasis [96, 106, 153, 168, 189]. Datasets derived from Internet connectivity will exclude a significant population, e.g., many argued that India’s mandatory COVID-19 contact tracing app excluded hundreds of millions due to access constraints, pointing to the futility of digital nation-wide tracing (also see [112]). Moreover, India’s data footprint is relatively new, being a recent entrant to 4G mobile data. Many respondents observed a bias towards upper-middle class problems, data, and deployments due to easier data access and revenue, as P8, CS/IS (computer science/information sciences researcher) put it, “rich people problems like cardiac disease and cancer, not poor people’s Tuberculosis, prioritised in AI [in India],” exacerbating inequities among those who benefit from AI and those who do not.

Several respondents pointed to missing data due to class, gender, and caste inequities in accessing and creating online content, e.g., many safety apps use data mapping to mark areas as unsafe, in order

---

3We are sympathetic to Dalit feminist scholars, like Rege, who have critiqued post-colonial or subaltern feminist thoughts for the lack of anti-caste scrutiny [162].

4Thanks to Murali Shanmugavelan for contributing these points.
to calculate an area-wide safety score for use by law enforcement [2, 9] (women’s safety is a pressing issue in India, in public consciousness since the 2012 Nirbhaya gang rape [75]. P4 (anti-caste, communications researcher) described how rape is socially (caste, culture, and religion) contextualised and some incidents get more visibility than others, in turn becoming data, in turn getting fed into safety apps—a perpetual source of unfairness. Many respondents were concerned that the safety apps were populated by middle-class users and tended to mark Dalit, Muslim, and slum areas as unsafe, potentially leading to hyper-patrolling in these areas.

Data was reported to be ‘missing’ due to artful user practices to manipulate algorithms, motivated by privacy, abuse, and reputation concerns. e.g., studies have shown how women users have ‘confused’ algorithms, motivated by privacy needs [130, 178]. Another class of user practices that happened outside of applications led to ‘off data’ traces. As an example, P17, CS/IS researcher, pointed to how auto rickshaw drivers created practices outside of ride-sharing apps, like calling passengers to verify landmarks (as Indian addresses are harder to specify [55]) or cancelling rides in-app (which used mobile payments) to carry out rides for a cash payment. Respondents described how data, like inferred gender, lacking an understanding of context was prone to inaccurate inferences.

Many respondents pointed to the frequent unavailability of socioeconomic and demographic datasets at national, state, and municipal levels for public fairness research. Some respondents reported on how the state and industry apparatus collected and retained valuable, large-scale data, but the datasets were not always made publicly available due to infrastructure and non-transparency issues. As P5, public policy researcher, described, “The country has the ability to collect large amounts of data, but there is no access to it, and not in a machine-readable format.” In particular, respondents shared how datasets featuring migration, incarceration, employment, or education, by sub-groups, were unavailable to the public. Scholarship like Bond’s caste report [184] argues that there is limited political will to collect and share socio-economic indicators by caste or religion.

A rich human infrastructure [182] from India’s public service delivery, e.g., frontline data workers, call-center operators, and administrative staff extends into AI data collection. However, they face disproportionate work burden, sometimes leading to data collection errors [37, 95, 143]. Many discussed how consent to a data worker stemmed from high interpersonal trust and holding them in high respect—relationships which may not be transitive to the downstream AI applications. In some cases, though, data workers have been shown to fudge data without actual conversations with affected people; efforts like jun sanwais (public hearings) and Janata Information Systems organized by the Mazdoor Kisan Shakti Sangathan are examples to secure representation through data. [97, 193].

**Mis-recorded identities** Statistical fairness makes a critical assumption so pervasively that it is rarely even stated: that user data corresponds one-to-one to people. However the one-to-one correspondence in datasets often fails in India. Ground truth on full names, location, contact details, biometrics, and their usage patterns can be unstable, especially for marginalised groups. User identity can be mis-recorded by the data collection instrument, assuming a static individual correspondence or expected behaviour. Since

---

| Sub-groups, Proxies and Harms |
|--------------------------------|
| **Caste** (17% Dalits, 8% Adivasi, 48% Other Backward Class (OBC)) [135] |
| • Societal harms: Human rights atrocities. Poverty, Land, knowledge and language battles [18, 92, 215]. |
| • Proxies: Surname. Skin tone. Occupation. Neighborhood. Language. |
| • Tech harms: Low literacy and phone ownership. Online misrepresentation & exclusion. Accuracy gap of Facial Recognition (FR). Limits of Fitzpatrick scale. Caste-based discrimination in tech [140]. |
| **Gender** (48.5% female) [47] |
| • Societal harms: Sexual crimes. Dowry. Violence. Female infanticide. |
| • Proxies: Name. Type of labor. Mobility from home. |
| • Tech harms: Accuracy gap in FR. Lower creditworthiness score. Recommendation algorithms favoring majority male users. Online abuse and ‘racily’ content issues. |
| **Religion** (80% Hindu, 14% Muslim, 6% Christians, Sikhs, Buddhists, Jains and indigenous) [135] |
| • Societal harms: Discrimination, lynching, vigilantism, and gang-rape against Muslims and others [11]. |
| • Proxies: Name. Neighborhood. Expenses. Work. Language. Clothing. |
| • Tech harms: Online stereotypes and hate speech, e.g., Islamophobia. Discriminatory inferences due to lifestyle, location, appearance. Targeted Internet disruptions. |
| **Ability** (5%-8% persons with disabilities) [163] |
| • Societal harms: Stigma. Inaccessible education, transport & work. |
| • Proxies: Non-normative facial features, speech patterns, body shape & movements. |
| • Use of assistive devices. |
| • Tech harms: Assumed homogeneity in physical, mental presentation. Paternalistic words and images. No accessibility mandate. |
| **Class** (30% live below poverty line, 48% on $2-$10/day) [160] |
| • Societal harms: Poverty. Inadequate food, shelter, health, & housing. |
| • Proxies: Spoken & written language(s). Mother tongue. Literacy. Feature / Smart Phone Ownership. Rural vs. urban. |
| • Tech harms: Linguistic bias towards mainstream languages. Model bias towards middle class users. Limited or lack of internet access. |
| **Gender Identity & Sexual Orientation** (No official LGBTQ+ data) |
| • Societal harms: Discrimination and abuse. Lack of acceptance and visibility, despite the recent decriminalisation [199]. |
| • Proxies: Gender declaration. Name. |
| • Tech harms: FR “outing” and accuracy. Gender binary surveillance systems (e.g., in dormitories). M/F ads targeting. Catfishing and extortion abuse attacks. |
| **Ethnicity** (4% NorthEast) [135] |
| • Societal harms: Racist slurs, discrimination, and physical attacks. |
| • Proxies: Skin tone. Facial features. Mother tongue. State. Name. |
| • Tech harms: Accuracy gap in FR. Online misrepresentation & exclusion. Inaccurate inferences due to lifestyle, e.g., migrant labor. |

---

| Table 1: Axes of potential ML (un)fairness in India |

---

conventional gender roles in Indian society lead to men having better access to devices, documentation, and mobility (see [64, 178]), women often borrowed phones. A few respondents pointed to how household dynamics impacted data collection, especially when using the door-to-door data collection method, e.g., how heads of households, typically men, often answered data-gathering surveys on behalf of women, but responses were recorded as women’s. Several AI-based applications use phone numbers as a proxy for identity in account creation (and content sharing) in the non-West, where mobile-first usage is common and e-mail is not [63]. However, this approach fails due to device sharing patterns [178], increased use of multiple SIM cards, and the frequency with which people change their numbers. Several respondents mentioned how location may not be permanent or tethered to a home, e.g., migrant workers regularly travel across porous nation-state boundaries.
Discriminated sub-groups and proxies AI systems in India remain under-analysed for biases, mirroring the limited public discourse on oppression. In Table 1, we present a summary of discriminated sub-groups in India, derived from our interviews and enriched through secondary research and statistics from authoritative sources, to substantiate attributes and proxies. Furthermore, we describe below some common discriminatory proxies and attributes that came up during our interviews. While the proxies may be similar to those in the West, the implementation and cultural logics may vary in India, e.g., P19, STS researcher, pointed to how Hijra community members (a marginalised intersex or transgender community) may live together in one housing unit and be seen as fraudulent or invisible to models using family units. Proxies may not generalise even within the country, e.g., asset ownership: “If you live in Mumbai, having a motorbike is a nuisance. If rural, you’re the richest in town.” (P9, CS/IS researcher).

- **Names**: are revelatory proxies for caste, religion, gender, and ethnicity, and have contributed to discrimination in India [15, 202], e.g., Banerjee or Khan connote to caste or religion location.
- **Zip codes**: can correspond to caste, religion, ethnicity, or class. Many Indian zip codes are heterogeneous, with slums and posh houses abutting in the same area [20, 36], unlike the rather homogeneous zip codes in the West from the segregated past [30].
- **Occupation**: traditional occupations may correspond to caste, gender, or religion; e.g., manual scavenging or butchery.
- **Expenditure**: on dietary and lifestyle items may be proxies for religion, caste, or ethnicity; e.g., expenses on pork or beef.
- **Skin tone**: may indicate caste, ethnicity, and class; however, unlike correlations between race and skin tone, correspondences to Indian sub-groups is weaker. Dark skin tones can be discriminated against in India [60]. Many respondents described how datasets under-collected darker skin tones and measurement scales like Fitzpatrick scale are not calibrated on diverse Indian skin tones.
- **Mobility**: can correlate to gender and disability. Indian women travel shorter distances and depend on public transport, for safety and cost [138]. Persons with disabilities often have lower mobility in India, due to a lack of ramps and accessible infrastructure [188].
- **Language**: Language can correspond to religion, class, ethnicity, and caste. Many AI systems serve in English, which only 10% of Indians understand [174]. India has 30 languages with over a million speakers. Everyday slurs such as ‘dhobi’, ‘kameena’, ‘pariah’, or ‘jungle’ are reported to be rampant online [15, 107].
- **Devices and infrastructure**: Internet access corresponds to gender, caste, and class, with 67% Internet users being males [102].
- **Documentation**: Several AI applications require state-issued documentation like Aadhaar or birth certificates, e.g., in finance. The economically poor are also reported to be document-poor [173].

4.1.2 Model considerations.

Over-fitting models to the privileged Respondents described how AI models in India overfitted to ‘good’ data profiles of the digitally-rich, privileged communities, as a result of poor cultural understanding and exclusion on part of AI makers. Respondents noted that the sub-groups that had access to underlying variables for data-rich profiles, like money, mobility, and literacy, were typically middle-class men. Model inputs in India appear to be disproportionately skewed due to large disparities in digital access. For instance, P11, tech policy researcher, illustrated how lending apps instantly determined creditworthiness through alternative credit histories built based on the user’s SMS messages, calls, and social networks (due to limited credit or banking history). Popular lending apps equate ‘good credit’ with whether the user called their parents daily, had stored over 58 contacts, played car-racing games, and could repay in 30 days [56]. Many respondents described how lending models imagined middle-class men as end-users—even with many microfinance studies showing that women have high loan repayment rates [69, 198]. In some cases, those with ‘poor’ data profiles subverted model predictions—as in P23’s (STS researcher) research on financial lending, where women overwhelmingly availed and paid back loans in the names of male relatives to avoid perceived gender bias in credit scoring. Model re-training left new room for bias, though, due to a lack of Fair-ML standards for India, e.g., an FR service used by police stations in eight Indian states retrained a western FR model on photos of Bollywood and regional film stars to mitigate the bias [61]—but Indian film stars are overwhelmingly fair-skinned, conventionally attractive, and able-bodied [108], not fully representative of the larger society.

Indic justice in models Popular fairness techniques, such as equal opportunity and equal odds, stem from epistemological and legal systems of the US (e.g., [62, 216]). India’s own justice approaches present new and relevant ways to operationalise algorithmic fairness locally. Nearly all respondents discussed reservations as a technique for algorithmic fairness in India. One of the restorative justice measures to repair centuries of subordination—reservations are a type of affirmative action enshrined in the Indian constitution [164]. Reservations assigns quotas for marginalised communities at the national and state levels.7 Originally designed for Dalits and Adivasis, reservations have expanded to include marginalised communities in the world’s most radical policies [25] (see [164] for more). Several studies have pointed to the successful redistribution of resources towards oppressed sub-groups, through reservations [43, 68, 152].

4.2 Double standards by ML makers

‘Bottom billion’ petri dishes Several respondents discussed how AI developers, both regional and international—private and state—treated Indian user communities as ‘petri dishes’ for models. Many criticised how neo-liberal AI (including those from the state) tended to treat Indians as ‘bottom billion data subjects’ in the periphery [211]—being subject to intrusive models, non-consensual automation, poor tech policies, inadequate user research, low-cost or free products that are low standard, and considered ‘unsaturated markets’. India’s diversity of languages, scripts, and populace has been

---

6National IDs are contested in the non-West, where they are used to deliver welfare in an ‘objective’ manner, but lead to populations falling through the cracks (see research on IDs in Aadhaar [161, 194], post-apartheid South Africa [65] and Sri Lankan Tamils and Côte d’Ivoirians [24]). Documentation has been known to be a weapon to dominate the non-literate in postcolonial societies [83]. Also see administrative violence [195].

7While the US Supreme court has banned various quotas [101], there is a history of quotas in the US, sometimes discriminatory, e.g., New Deal black worker quotas [58] and anti-Semitic quotas in universities [88]. Quotas in India are legal and common. Thanks to Martin Wattenberg for this point.
reported to be attractive for improving model robustness and training data [14]. Many discussed how low quality designs, algorithms, and support were deployed for Indians, attributing to weak tech policy and enforcement of accountability in India. Several respondents described how AI makers had a transactional mindset towards Indians, seeing them as agency-less data subjects that generated large-scale behavioural traces to improve models.

In contrast to how AI industry and research were moderately responsive to user bias reports in the West, recourse and redress for Indians were perceived to be non-existent. Respondents described that when recourse existed, it was often culturally-insensitive or dehumanising, e.g., a respondent was violently questioned about their clothing by staff of a ride-sharing application, during redressal for an assault faced in a taxi (also see [177] for poor recourse). Several respondents described how lack of recourse was even more dire for marginalised users, e.g., P14 (CS/IS researcher) described, “[Ordering a ride-share] a person with a disability would choose electronic payment, but the driver insisted on cash. They said they are blind and wanted to pay electronically, but the driver declined and just moved on. No way to report it.” Even when feedback mechanisms were included, respondents shared that they were not always localised for India, and incidents were not always recognised unless an activist contacted the company staff directly. Many respondents shared how street-level bureaucrats, administrative offices, and front line workers—the human infrastructures [182] who played a crucial role through village-level audits of documents, e.g., 8eral respondents discussed how Indian AI engineers were largely “clean faeces in the morning and now she is a guinea pig for new AI.”

The human infrastructures [182] who played a crucial role through village-level audits of documents, e.g., 8eral respondents discussed how Indian AI engineers were largely “clean faeces in the morning and now she is a guinea pig for new AI.”

In high-tech solutions, combined with limited technology policy fervent rhetoric; whereas in practice, accuracy and performance of high-stakes solutions, e.g., how predictive policing was reported to have very low accuracy and failed to distinguish between boy and girl children [5]. Some respondents mentioned how a few automation solutions were announced following public sentiment, but turned into surveillance e.g., how predictive policing in Delhi and FR in train stations was announced after Nirbhaya’s gruesome gangrape in 2012 and women’s safety incidents [28, 33]. Several respondents pointed to how automation solutions had fervent rhetoric; whereas in practice, accuracy and performance of systems were low. Many described how disproportionate confidence in high-tech solutions, combined with limited technology policy engagement among decision-makers, appeared to lead to sub-par high-stakes solutions, e.g., the FR service used by Delhi Police was reported to have very low accuracy and failed to distinguish between boy and girl children [5]. Some respondents mentioned how a few automation solutions were announced following public sentiment, but turned into surveillance e.g., how predictive policing in Delhi and FR in train stations was announced after Nirbhaya’s gruesome gangrape in 2012 and women’s safety incidents [28, 33].

Respondents described how public sector AI projects in India were viewed as modernising efforts to overcome age-old inefficiencies in resource delivery (also in [175]). The AI embrace was attributed to follow the trajectory of recent high-tech interventions (such as Aadhaar, MGNREGA payments, and the National Register of Citizens (NRC)). Researchers have pointed to the aspirational role played by technology in India, signifying symbolic meanings of modernity and progress via technocracy [149, 150, 176]. AI for societal benefit is a pivotal development thrust in India, with a focus on healthcare, agriculture, education, smart cities, and mobility [1]—influencing citizen imaginaries of AI. In an international AI perceptions survey (2019), Indians ranked first in rating AI as ‘exciting’, ‘futuristic’ and ‘mostly good for society’ [110].

Several respondents pointed to how Dalit and Muslim bodies were used as test subjects for AI surveillance, e.g., pointing to how human efficiency trackers were increasingly deployed among Dalit sanitation workers in cities like Panchkula and Nagpur. Equipped with microphones, GPS, cameras, and a SIM, the trackers allowed detailed surveillance of movement and work, leading to some women workers avoiding restrooms for fear of camera capture, avoiding sensitive conversations for fear of snooping, and waiting for the tracker to die before going home [111]. Such interventions were criticised for placing power in the hands of dominant-caste supervisors. P21 (legal researcher) pointed out that surveillance has historically been targeted at the working-poor, “the house cleaner who is constantly suspected of stealing dried fruits or jewellery. Stepping out of their house means that their every move is tracked. Someone recording the face, the heartbeat..under the pretext of efficiency. Her job is to clean faeces in the morning and now she is a guinea pig for new AI.”

Entrenched privilege and distance Nearly all respondents described how Indian AI engineers were largely privileged class and caste males.9 For e.g., P17 (CS/IS researcher) described, “Who is designing AI? Incredibly entitled, Brahmin, certainly male. They’ve never encountered discrimination in their life. These guys are talking about primitive women. If they’re designing AI, they haven’t got a clue about the rest of the people. Then it becomes fairness for who?” Many respondents described how the Indian technology sector claimed to be ‘merit-based’, open to anyone highly gifted in the technical sciences; but many have pointed to how merit is a function of caste privilege [196, 207]. Many, though not all, graduates of Indian Institutes of Technology, founders of pioneering Indian software companies and nearly all Nobel prize winners of Indian origin have come from privileged castes and class backgrounds [71, 196]. As P21 (legal researcher) explained the pervasive privilege in AI, “Silicon Valley Brahmins [Indians] are not questioning the social structure they grew up in, and white tech workers do not understand caste to spot and mitigate obvious harms.” While engineers and researchers are mostly privileged everywhere, the stark socio-economic disparities between Indian engineers and the marginalised communities may further amplify the distances.

4.3 Unquestioning AI aspiration

AI euphoria Several respondents described how strong aspiration for AI for socio-economic upliftment was accompanied by high trust in automation, limited transparency, and the lack of an empowered Fair-ML ecosystem in India. Contrast with the West, where a large, active stakeholder ecosystem (of civil society, journalists, and law makers) is AI-literate and has access to open APIs and data. Many respondents described how public sector AI projects in India were viewed as modernising efforts to overcome age-old inefficiencies in resource delivery (also in [175]). The AI embrace was attributed to follow the trajectory of recent high-tech interventions (such as Aadhaar, MGNREGA payments, and the National Register of Citizens (NRC)). Researchers have pointed to the aspirational role played by technology in India, signifying symbolic meanings of modernity and progress via technocracy [149, 150, 176]. AI for societal benefit is a pivotal development thrust in India, with a focus on healthcare, agriculture, education, smart cities, and mobility [1]—influencing citizen imaginaries of AI. In an international AI perceptions survey (2019), Indians ranked first in rating AI as ‘exciting’, ‘futuristic’ and ‘mostly good for society’ [110].

Several respondents pointed to how automation solutions had fervent rhetoric; whereas in practice, accuracy and performance of systems were low. Many described how disproportionate confidence in high-tech solutions, combined with limited technology policy engagement among decision-makers, appeared to lead to sub-par high-stakes solutions, e.g., the FR service used by Delhi Police was reported to have very low accuracy and failed to distinguish between boy and girl children [5]. Some respondents mentioned how a few automation solutions were announced following public sentiment, but turned into surveillance e.g., how predictive policing in Delhi and FR in train stations was announced after Nirbhaya’s gruesome gangrape in 2012 and women’s safety incidents [28, 33].

9India fares slightly better than the US in gender representation in the tech workforce; however, gender roles and safety concerns lead to nearly 80% of women leaving computing by their thirties (coinciding with family/parenting responsibilities) [200].

9A confidence threshold of 80-95% is recommended for law enforcement AI [54]
Disputing AI4All Many respondents pointed to how emerging ‘good’ deployments tended to leave out minorities. e.g., P29 (LGBTQ+ activist) discussed how AI was justified in the public domain, e.g., surveillance for smart cities,11 as women’s safety measures, but tended to invisibilise transgender members or increase monitoring of Dalit and Muslim areas, e.g., a FR was deployed outside women’s restrooms to detect intrusion by non-female entrants, potentially leading to violence against transgender members.

Many respondents expressed concern over AI advances in detecting sexuality, criminality, or terrorism (e.g., [187, 212]) potentially being exported to India and harming minorities. P29 remarked on targeted attacks [26], “part of the smart cities project is a Facial Recognition database where anyone can upload images. Imagine the vigilantism against dalit, poor, Muslims, trans persons if someone uploads a photo of them and it was used for sex offenders [arrests].”

Algorithmic opacity and authority In contrast to the ‘black box AI problem’, i.e., even the humans who design models do not always understand how variables are being combined to make inferences [171], many respondents discussed an end-to-end opacity of inputs, model behaviour, and outcomes in India. Fairness in India was reported to suffer from a lack of access to contributing datasets, APIs, and documentation, with several respondents describing how challenging it was for researchers and civil society to assess the high-stakes AI systems. As P11 described, “Opacity is quite acute [in India]. People talk about blackboxes, reverse engineering inputs from outputs. What happens when you don’t have the output? What happens when you can’t reverse engineer at all?”. AI’s ‘neutral’ and ‘human-free’ associations lent credence to its algorithmic authority. In January 2020, over a thousand protestors were arrested during protests in Delhi, aided by FR. The official statement was, “This is a software. It does not see faith. It does not see clothes. It only sees the face and through the face the person is caught.” [5]. While algorithms may not be trained on sub-group identification, proxies may correspond to Dalits, Adivasis, and Muslims disproportionately. e.g., according to the National Crime Records Bureau (NCRB) in 2015, 34% of undertrials were Dalits and Adivasis (25% of the population); 20% were Muslims (14% of population); and 70% were non-literate (26% of the population) [203].

Several respondents discussed a lack of inclusion of diverse stakeholders in decision-making processes, laws, and policies for public sector AI. Some talked about a colonial mindset of tight control in decision-making on automation laws, leading to repressive and monoscopic views by the judiciary and state. P5 (public policy researcher) pointed to how mission and vision statements for public sector AI tended to portray AI like magic, rather than contending with the realities of “how things worked on-the-ground in a developing country”. Additionally, respondents pointed to significant scope creep in high-stakes AI, e.g., a few mentioned how the tender for an FR system was initially motivated by detecting criminals, later missing children, to then arresting protestors [5].

Questioning AI power Algorithmic fairness requires a buffer zone of journalists, activists, and researchers to keep AI system builders accountable. Many respondents described how limited debate and analysis of AI in India led to a weak implementation of Fair-ML in India. Issues of algorithmic bias were not widely raised in the public consciousness in India, at the time of the study. Respondents described how technology journalists in India—a keystone species for public discourse—covered app launches and investments, and less on algorithms and their impacts. P15 (journalist) pointed out that journalists may face disapproval for questioning certain narratives. “The broader issue is that AI biases do not come up in Indian press. Our definition of tech journalism has been about covering the business of tech [...] It is different from the US, where there is a more combative relationship. We don’t ask tough questions of the state or tech companies.” A seminal report by NewsLaundry/Oxfam described how privileged castes comprised Indian news media, invisibilising the vast oppressed majority in public discourse [145].

5 TOWARDS AI FAIRNESS IN INDIA

In Section 4, we demonstrated that several underlying assumptions about algorithmic fairness and its enabling infrastructures fail in the Indian context. To meaningfully operationalise algorithmic fairness in India, we extend the spirit of Dr. B.R. Ambedkar’s call to action, “there cannot be political reform without social reform” [18]. We need to substantively innovate on how to empower oppressed communities and enable justice within the surrounding socio-political infrastructures in India. Missing Indic factors and values in data and models, compounded by double standards and disjuncture in ML, deployed in an environment of unquestioning AI aspiration, face the risk of reinforcing injustices and harms. We need to understand and design for end-to-end chains of algorithmic power, including how AI systems are conceived, produced, utilised, appropriated, assessed, and contested in India. We humbly submit that these are large, open-ended challenges that have perhaps not received much focus or are considered large in scope. However, a Fair-ML strategy for India needs to reflect its deeply plural, complex, and contradictory nature, and needs to go beyond model fairness. We propose an AI fairness research agenda in India, where we call for action along three critical and contingent pathways: Recontextualising, Empowering, and Enabling. Figure 2 shows the core considerations of these pathways. The pathways present opportunities for cross-disciplinary and cross-institutional collaboration. While it is important to incorporate Indian concepts of fairness into AI that impacts Indian communities, the broader goal should be to develop general approaches to fairness that reflect the needs of local communities and are appropriate to the local contexts.

5.1 Recontextualising Data and Models

How might existing algorithmic fairness evaluation and mitigation approaches be recontextualised to the Indian context, and what novel challenges does it give rise to?

5.1.1 Data considerations.

Data plays a critical role in measurements and mitigations of algorithmic bias. However, as seen in Section 4, social, economic, and infrastructural factors challenge the reliance on datasets in India. Based on our findings, we outline some recommendations and put forth critical research questions regarding data and its uses in India.

Due to the challenges to completeness and representation discussed in Section 4, we must be (even more than usual) sceptical of Indian datasets until they are trust-worthy. For instance, how could fairness research handle the known data distortions guided by

11 Indian cities are smart city sites, to be equipped with intelligent traffic, waste and energy management, and CCTV crime monitoring. http://smartcities.gov.in/content.
Figure 2: Research pathways for Algorithmic Fairness in India.

traditional gender roles? What are the risks in identifying caste in models? Should instances where models are deliberately ‘confused’ (see Section 4) be identified, and if so what should we do with such data? We must also account for data voids [78] for which statistical extrapolations might be invalid.

The vibrant role played by human infrastructures in providing, negotiating, collecting, and stewarding data points to new ways of looking at data as dialogue, rather than operations. Such data gathering via community relationships lead us to view data records as products of both beholder and the beheld. Building ties with community workers in the AI dataset collection process can be a starting point in creating high quality data, while respecting their situated knowledge. Combining observational research and dataset analysis will help us avoid misreadings of data. Normative frameworks (e.g., perhaps ethics of care [22, 87, 217]) may be relevant to take into account these social relations. A related question is how data consent might work fairly in India. One approach could be to create transitive informed consent, built upon personal relationships and trust in data workers, with transparency on potential downstream applications. Ideas like collective consent [172], data trusts [141], and data co-ops may enhance community agency in datasets, while simultaneously increasing data reliability.

Finally, at a fundamental level, we must question the categories and constructs we model in datasets, and how we measure them. As well as the situatedness of social categories such as gender (cf. Hijra) and race [85], ontologies of affect (sentiment, inappropriateness, etc.), taboos (Halal, revealing clothing, etc.), and social behaviours (headshakes, headwear, clothing, etc) are similarly contextual. How do we justify the “knowing” of social information by encoding it in data? We must also question if being “data-driven” is inconsistent with local values, goals and contexts. When data are appropriate for endemic goals (e.g., caste membership for quotas), what form should their distributions and annotations take? Linguistically and culturally pluralistic communities should be given voices in these negotiations in ways that respect Indian norms of representation.

5.1.2 Model and model (un)fairness considerations.

The prominent axes of historical injustices in India listed in Table 1 could be a starting point to detect and mitigate unfairness issues in trained models (e.g., [42, 218]), alongside testing strategies, e.g. perturbation testing [156], data augmentation [220], adversarial testing [117], and adherence to terminology guidelines for oppressed groups, such as SIGACCESS. However, it is important to note that operationalising fairness approaches from the West to these axes is often nontrivial. For instance, personal names act as a signifier for various socio-demographic attributes in India, however there are no large datasets of Indian names (like the US Census data, or the SSA data) that are readily available for fairness evaluations. In addition, disparities along the same axes may manifest very differently in India. For instance, gender disparities in the Indian workforce follow significantly different patterns compared to the West. How would an algorithm made fairer along gender based on datasets from the West be decontextualised and recontextualised for India?

Another important consideration is how the algorithmic fairness interventions work with the existing infrastructures in India that surrounds decision making processes. For instance, how do they work in the context of established fairness initiatives such as reservations/quotas? As an illustration, compared to the US undergraduate admission process of selection from a pool of candidates, the undergraduate admissions in India is done through various joint seat allocation processes, over hundreds of programmes, across dozens of universities and colleges that takes into account test scores, ordered preference lists provided by students, as well as various affirmative action quotas [31]. The quota system gives rise to the problem of matching under distributional constraints, a known problem in economics [23, 79, 105], but has not received attention within the FAccT community (although [51] is related). First-order Fair-ML problems could include representational biases of caste and other sub-groups in NLP models, biases in Indic language NLP including challenges from code-mixing, Indian subgroup biases in computer vision, tackling online misinformation, benchmarking using Indic datasets, and fair allocation models in public welfare.

5.2 Empowering Communities

Recontextualising data and models can only go so far without participatorily empowering communities in identifying problems, specifying fairness expectations, and designing systems.
5.2.1 Participatory Fair-ML knowledge systems. Context-free assumptions in Fair-ML, whether in homegrown or international AI systems, can not just fail, but produce harm inadvertently when applied to different infrastructural and cultural systems. As Mbembe describes the Western epistemic tradition, the knowing subject is enclosed in itself and produces supposedly objective knowledge of the world, “without being part of that world, and he or she is by all accounts able to produce knowledge that is supposed to be universal and independent of context” [131]. How can we move away from the ‘common good’ defined by the researcher—the supposedly all-knowing entity who has the expertise and experience necessary to identify what is of benefit to all [144]? Humbly creating grassroots commitments with vulnerable communities is an important first step. Our study discusses how caste, religion, and tribe are eluded even within the Indian technology discourse and policy. Modes of epistemic production in Fair-ML should enable marginalised communities to produce knowledge about themselves in the policies or designs. Grassroots efforts like Deep Learning Indaba [3] and Khipu [7] are exemplar of bootstrapping AI research in communities. Initiatives like Design Beku [4] and SEWA [10] are excellent decolonial examples of participatory co-designing with under-served communities.

5.2.2 Low-resource considerations. India’s heterogeneous literacies, economies, and infrastructures mean that Fair-ML researchers’ commitment should go beyond model outputs, to deployments in accessible systems. Half the population of India is not online. Layers of the stack like interfaces, devices, connectivity, languages, and costs are important to ensure access. Learning from computing fields where constraints have been embraced as design material like ICTD [204] and HCD [67] can help, e.g., via delay-tolerant connectivity, low cost devices, text-free interfaces, intermediaries, and NGO partnerships (see [45, 74, 86, 115, 132, 179, 209]). Data infrastructures to build localised datasets would enhance access equity (e.g., [8, 181]).

5.2.3 First-world care in deployments. Critiques were raised in our study on how neo-liberal AI followed a broader pattern of extraction from the ‘bottom billion’ data subjects and labourers. Low costs, large and diverse populations, and policy infirmities have been cited as reasons for following double standards in India, e.g., in non-consensual human trials and waste dumping [123] (also see [137]). Past disasters in India, like the fatal Union Carbide gas leak in 1984—one of the world’s worst industrial accidents—point to faulty design and low quality, double standards for the ‘third world’ [206]. Similarly, unequal standards, inadequate safeguards, and dubious applications of AI in the non-West can lead to catastrophic effects (similar analogies have been made for content moderation [165, 177]). Fair-ML researchers should understand the systems into which they are embedding, engage with Indian realities and user feedback, and whether the recourse is meaningful.

5.3 Enabling Fair-ML Ecosystems. AI is increasingly perceived as a masculine, hype-filled, technoutopian enterprise, with nations turning into AI superpowers [170]. AI is even more aspirational and consequential in non-Western nations, where it performs an ‘efficiency’ function in distributing scarce socio-economic resources and differentiating economies. For Fair-ML research to be impactful and sustainable, it is crucial for researchers to enable a critically conscious Fair-ML ecosystem.

5.3.1 Ecosystems for accountability. Bootstrapping an ecosystem made up of civil society, media, industry, judiciary, and the state is necessary for accountability in Fair-ML (recall the US FR example). Moving from ivory tower research approaches to solidarity with various stakeholders through partnerships, evidence-based policy, and policy maker education can help create a sustainable Fair-ML ecosystem based on sound empirical and ethical norms, e.g., we should consider research with algorithmic advocacy groups like Internet Freedom Foundation [6], that have advanced landmark changes in net neutrality and privacy. Efforts like the AI Observatory to catalogue, understand harms, and demand accountability of automated decision support systems in India are crucial first steps [100]. Technology journalism is a keystone of equitable automation and needs to be fostered for AI.

5.3.2 Critical transparency. Incrustability suppresses algorithmic fairness. Besides the role played by ecosystem-wide regulations and standards, radical transparency should be espoused and enacted by Fair-ML researchers committed to India. Transparency on datasets, processes, and models (e.g., [34, 77, 136]), openly discussing limitations, failure cases, and lessons learnt can help move from the ‘magic pill’ role of fairness as a checklist for ethical issues in India—to a more pragmatic, flawed, and evolving scientific function.

6 CONCLUSION. As AI becomes global, algorithmic fairness naturally follows. Context matters. We must take care to not copy-paste the western-normative fairness everywhere. We presented a qualitative study and discourse analysis of algorithmic power in India, and found that algorithmic fairness assumptions are challenged in the Indian context. We found that data was not always reliable due to socio-economic factors, ML products for Indian users suffer from double standards, and AI was seen with unquestioning aspiration. We called for an end-to-end re-imagining of algorithmic fairness that involves re-contextualising data and models, empowering oppressed communities, and enabling fairness ecosystems. The considerations we identified are certainly not limited to India; likewise, we call for inclusively evolving global approaches to Fair-ML.

7 ACKNOWLEDGEMENTS. Our thanks to the experts who shared their knowledge and wisdom with us: A. Aneesh, Aishwarya Lakshmiratan, Ameen Jauhar, Amit Sethi, Anil Joshi, Arindrajit Basu, Avinash Kumar, Chiranjeeb Bhattacharya, Dhruv Lakra, George Sebastian, Jacki O Neill, Mainack Mondal, Maya Indira Ganesh, Murali Shanmugavelan, Nandana Sengupta, Neha Kumar, Rahul De, Rahul Mathan, Rajesh Veeraraghavan, Ranjit Singh, Ryan Joseph Figueiredo (Equal Asia Foundation), Savita Bailur, Sayomdebl Mukherjee, Shanti Raghavan, Shyam Suri, Smita, Sriram Somanchi, Surya Yengde, Vidushi Marda, and Vivek Srinivasan, and others who wish to stay anonymous. To Murali Shanmugavelan for educating us and connecting this paper to anti-caste emancipatory politics and theories. To Jose M. Faleiro, Daniel Russell, Jess Holbrook, Fernanda Viegas, Martin Wattenberg, Alex Hanna, and Reena Jana for their invaluable feedback.
REFERENCES

[1] National Strategy for Artificial Intelligence (SAIALL). Niti Aayog. 2020. https://www.niti.gov.in/sites/default/files/2020/09/05/niti-sai-all-strategy-tech-to-identify-1000-individuals-at-a-recent-riot.-Accepted-on-07/28/2020.

[2] Internet Freedom Foundation. https://internetfreedom.in/. 2020. (Accessed on 08/29/2020).

[3] SEWA. http://www.sewa.org/. 2019. (Accessed on 08/29/2020).

[4] Abhishek Baxi. 2018. Law Enforcement Agencies in India Are Using Artificial Intelligence in India. https://www.deeplearningindaba.com/2020/08/16/. (Accessed on 08/16/2020).

[5] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2017. Fairness in machine learning. Cambridge Quarterly of Healthcare Ethics 26, 2 (2017), 467–479.

[6] Abeba Birhane. 2020. Algorithmic colonization of Africa. SCRIPTed 17 (2020), 89.

[7] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI ’19). Montreal, QC, Canada, ACM, 1–10.

[8] Eric Brewer, Michael Demmer, Bowei Du, Melissa Ho, Matthew Kan, Sergiu Nedevschi, Joyojeet Pal, Rabin Patra, Sonesh Surana, and Kevin Fall. 2005. The case for technology in developing regions. Computer 38, 5 (2005), 25–38.

[9] Kayla K. McBauliffe, Joichi Ito, and Surya Mattu. 2016. Counting cows, not rural health indicators. In Proceedings of the 2016 Conference on Fairness, Accountability and Transparency (FAccT ’16). Montreal, QC, Canada, ACM, 1–10.

[10] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias — ProPublica. https://www.propublica.org/article/machine-bias-not-slang-but-community-where-i-belong-assam-womans-online-job-application-rejected-due-to-surname-625556. (Accessed on 09/28/2020).

[11]ạm 2019. Labour market discrimination in Delhi: Evidence from a field experiment. Journal of Comparative Economics 37, 1 (2009), 14–27.

[12] Anna Clark. 2013. ZIP Code History: How They Define Us | The New Republic. https://newrepublic.com/article/112558/zip-code-history-how-they-define-us. (Accessed on 09/24/2020).

[13] Kate Crawford. 2013. The hidden biases in big data. Foreign Policy. 149–155.

[14] Uma Chakravarti. 1993. Conceptualising Brahmanical patriarchy in early India: Gender, caste, and class status. Economic and Political Weekly 28, 6 (2003), 25–38.

[15] Eric Brewer, Michael Demmer, Bowei Du, Melissa Ho, Matthew Kan, Sergiu Nedevschi, Joyojeet Pal, Rabin Patra, Sonesh Surana, and Kevin Fall. 2005. The case for technology in developing regions. Computer 38, 5 (2005), 25–38.

[16] Andrew Cotter, Heinrich Jiang, Maya R Gupta, Serena Wang, Taman Narayan, Seunghl You, and Karthik Sridharan. 2019. Optimization with Non-Differentiable Constraints with Applications to Fairness, Recall, Churn, and Other Goals. Journal of Machine Learning Research 20, 172 (2019), 1–59.

[17] Kate Crawford. 2013. The hidden biases in big data. Harvard business review 1, 1 (2013), 841.

[18] Kate Crawford. 2013. Think again: Big data. Foreign Policy 149–155.

[19] Anubha Bhonsle and Pallavi Prasad. 2020. Counting cows, not rural health indicators. In Proceedings of the 2020 Conference on Fairness, Accountability and Transparency (FAccT ’20). March 1–10, 2020, Virtual Event, Canada.
