# 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, but pay less attention to what is fair in local contexts.

Conventional measurements of algorithmic fairness make several assumptions based on Western institutions and infrastructures. To illustrate, consider Facial Recognition (FR), where demonstration

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 corresponds 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 services 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.,* the 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 Rawl'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 Marathi) 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 hierarchical 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.[3] 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. A decolonial perspective (borrowed from Latin American and African scholars like [66, 70, 72, 134, 210]) helps us confront inequalities from colonisation in India, providing new openings for knowledge and justice in AI fairness research. To Dr. B. R. Ambedkar and Periyar E. V. Ramasamy, colonialism predates the British era, and decolonisation is a continuum. For Dalit emancipatory politics, deconstructing colonial ideologies of the powerful, superior, and privileged begins by removing influences and privileges of dominant-caste members.[4]

**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 underserved communities and technology, with backgrounds in HCI,

---

[3]We are sympathetic to Dalit feminist scholars, like Rege, who have critiqued postcolonial or subaltern feminist thoughts for the lack of anti-caste scrutiny [162]
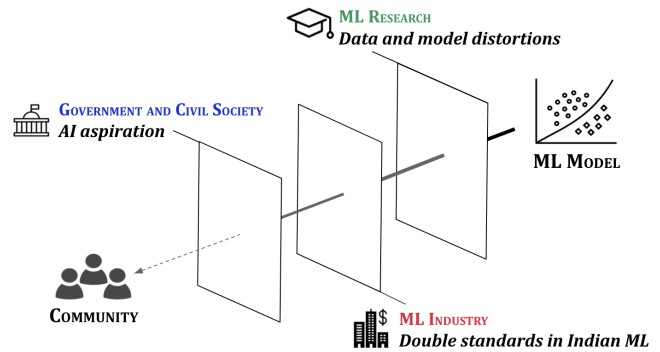[4]Thanks to Murali Shanmugavelan for contributing these points.



**Figure 1: Algorithmic powerful in India, where the distance between models and oppressed communities is large**

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

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 socio-economic and demographic datasets at national, state, and municipal levels for public fairness research. Some respondents reported on how the state and industry apparati 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 Sangatan 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.[5] 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

---

[5]This issue is somewhat related to what [148] call "Non-individual agents".

### Sub-groups, Proxies and Harms

**Caste** (17% Dalits; 8% Adivasi; 40% 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 'racey' content issues. Low Internet access.

**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 [50].
- *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 'junglee' 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].[6]

#### 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 other backward castes, women, persons with disabilities, and religious minorities. Depending on the policy, reservations can allocate quotas from 30% to 80%. Reservations in India have been described as one of 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

---

[6]National 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].

[7]While 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 in providing recourse to marginalised Indian communities—were removed in AI systems. Further, the high-tech illegibility of AI was noted to render recourse out of reach for groups marginalised by literacy, legal, and educational capital (see [208] for 'hiding behind a computer'). As P12 (STS researcher) explained, "*If decisions are made by a centralised server, communities don't even know what has gone wrong, why [welfare] has stopped, they don't know who to go to to fix the problem.*" Many described how social audits and working with civil society created a better understanding and accountability.[8]

Some 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 AI makers and researchers, including regional makers, were heavily distant from the Indian communities they served. Several respondents discussed 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].[10] 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].

---

[8]Social audits like jan sanwais have long gauged effectiveness of civil programmes through village-level audits of documents, *e.g.,* to curb corrupt funds siphoning [154].

[9]India 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].
[10]A confidence threshold of 80-95% is recommended for law enforcement AI [54]

**Disputing AI4All** Many respondents pointed to how emerging '4good' 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 reticent 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

---

[11] 98 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.
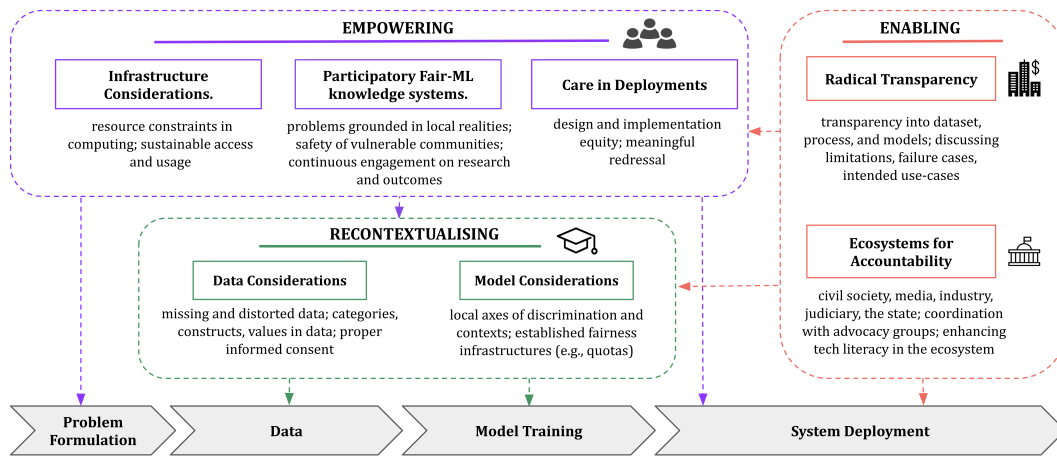
Figure 2: Research pathways for Algorithmic Fairness in India.

by traditional gender roles? What are the dangers of 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 participatorily co-designing with under-served communities.

### 5.2.2 Low-resource considerations.

India's heterogeneous literacies, economics, 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 HCI4D [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, techno-utopian 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.

Inscrutability 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 sufffer 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

# REFERENCES

[1] 2018. *National Strategy for Artificial Intelligence #AI4ALL*. Niti Aayog.
[2] 2020. *Citizen COP Foundation*. https://www.citizencop.org
[3] 2020. *Deep Learning Indaba*. https://deeplearningindaba.com/2020/
[4] 2020. *Design Beku*. https://designbeku.in/
[5] 2020. India used facial recognition tech to identify 1,100 individuals at a recent riot | TechCrunch. https://techcrunch.com/2020/03/11/india-used-facial-recognition-tech-to-identify-1100-individuals-at-a-recent-riot. (Accessed on 07/28/2020).
[6] 2020. *Internet Freedom Foundation*. https://internetfreedom.in/
[7] 2020. *Khipu AI*. https://github.com/khipu-ai
[8] 2020. *Lacuna Fund*. https://lacunafund.org/
[9] 2020. *Safetipin*. https://safetipin.com/
[10] 2020. *SEWA*. http://www.sewa.org/
[11] Delna Abraham and Ojaswi Rao. [n.d.]. 84% Dead In Cow-Related Violence Since 2010 Are Muslim; 97% Attacks After 2014 | IndiaSpend. https://archive.indiaspend.com/cover-story/86-dead-in-cow-related-violence-since-2010-are-muslim-97-attacks-after-2014-2014. (Accessed on 08/16/2020).
[12] Oshin Agarwal, Yinfei Yang, Byron C Wallace, and Ani Nenkova. 2020. Entity-Switched Datasets: An Approach to Auditing the In-Domain Robustness of Named Entity Recognition Models. *arXiv preprint arXiv:2004.04123* (2020).
[13] Digital Government Agency (Agesic). 2019. Artificial Intelligence for the digital government | English version. https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/sites/agencia-gobierno-electronico-sociedad-informacion-conocimiento/files/documentos/publicaciones/IA%20Strategy%20-20english%20version.pdf. In *AI whitepaper*.
[14] Varun Aggarwal. 2018. India's mess of complexity is just what AI needs | MIT Technology Review. https://www.technologyreview.com/2018/06/27/240474/indias-mess-of-complexity-is-just-what-ai-needs/. (Accessed on 09/18/2020).
[15] Saumya Agrawal. 2020. Chutia| 'Chutia not slang, but community where I belong': Assam woman's online job application rejected due to surname | Trending & Viral News. https://www.timesnownews.com/the-buzz/article/chutia-not-slang-but-community-where-i-belong-assam-womans-online-job-application-rejected-due-to-surname/625556. (Accessed on 09/28/2020).
[16] Amazon. 2020. We are implementing a one-year moratorium on police use of Rekognition. https://blog.aboutamazon.com/policy/we-are-implementing-a-one-year-moratorium-on-police-use-of-rekognition. (Accessed on 08/29/2020).
[17] BR Ambedkar. 1916. Castes in India: Their mechanism, genesis and development (Vol. 1). *Columbia: Indian Antiquary. Ambedkar, BR (1936). Annihilation of Caste. Jullundur: Bheem Patrika Publications* (1916).
[18] Bhimrao Ramji Ambedkar. 2014. *Annihilation of caste: The annotated critical edition*. Verso Books.
[19] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias — ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. (Accessed on 07/30/2020).
[20] Arjun Appadurai. 2000. Spectral housing and urban cleansing: notes on millennial Mumbai. *Public culture* 12, 3 (2000), 627–651.
[21] Payal Arora. 2016. Bottom of the data pyramid: Big data and the global south. *International Journal of Communication* 10 (2016), 19.
[22] Peter M Asaro. 2019. AI ethics in predictive policing: From models of threat to an ethics of care. *IEEE Technology and Society Magazine* 38, 2 (2019), 40–53.
[23] Itai Ashlagi, Amin Saberi, and Ali Shameli. 2020. Assignment mechanisms under distributional constraints. *Operations Research* 68, 2 (2020), 467–479.
[24] Savita Bailur, Devina Srivastava, and Hélène (Caribou Digital) Smertnik. 2019. Women and ID in a digital age: Five fundamental barriers and new design questions. https://savitabailur.com/2019/09/09/women-and-id-in-a-digital-age-five-fundamental-barriers-and-new-design-questions/. (Accessed on 08/02/2020).
[25] Robert Baker. 2001. Bioethics and Human Rights: A Historical Perspective. *Cambridge Quarterly of Healthcare Ethics* 10, 3 (2001), 241–252. https://doi.org/10.1017/S0963180101003048
[26] Shakuntala Banaji and Ram Bhat. 2019. *WhatsApp Vigilantes: An exploration of citizen reception and circulation of WhatsApp misinformation linked to mob violence in India*. Department of Media and Communications, LSE.
[27] Abhijit Banerjee, Marianne Bertrand, Saugato Datta, and Sendhil Mullainathan. 2009. Labor market discrimination in Delhi: Evidence from a field experiment. *Journal of comparative Economics* 37, 1 (2009), 14–27.
[28] Soumyarendra Barik. 2020. Facial recognition based surveillance systems to be installed at 983 railway stations across India. https://www.medianama.com/2020/01/223-facial-recognition-system-indian-railways-facial-recognition/. (Accessed on 10/03/2020).
[29] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2017. Fairness in machine learning. *NIPS Tutorial* 1 (2017).
[30] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.
[31] Surender Baswana, Partha Pratim Chakrabarti, Sharat Chandran, Yashodhan Kanoria, and Utkarsh Patange. 2019. Centralized admissions for engineering colleges in India. *INFORMS Journal on Applied Analytics* 49, 5 (2019), 338–354.
[32] Abhishek Baxi. 2018. Law Enforcement Agencies In India Are Using Artificial Intelligence To Nab Criminals. https://www.forbes.com/sites/baxiabhishek/2018/09/28/law-enforcement-agencies-in-india-are-using-artificial-intelligence-to-nab-criminals-heres-how. (Accessed on 08/30/2020).

[33] BBC. 2019. Nirbhaya case: Four Indian men executed for 2012 Delhi bus rape and murder - BBC News. https://www.bbc.com/news/world-asia-india-51969961. (Accessed on 09/01/2020).
[34] Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604.
[35] Andre Beteille. 1990. Race, caste and gender. *Man* (1990), 489–504.
[36] Naveen Bharathi, Deepak V Malghan, and Andaleeb Rahman. 2018. Isolated by caste: Neighbourhood-scale residential segregation in Indian metros. *IIM Bangalore Research Paper* 572 (2018).
[37] Anubha Bhonsle and Pallavi Prasad. 2020. Counting cows, not rural health indicators. https://ruralindiaonline.org/articles/counting-cows-not-rural-health-indicators/. (Accessed on 08/02/2020).
[38] Reuben Binns. 2018. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*. 149–159.
[39] Abeba Birhane. 2020. Algorithmic colonization of Africa. *SCRIPTed* 17 (2020), 389.
[40] PR Blake, K McAuliffe, J Corbit, TC Callaghan, O Barry, A Bowie, L Kleutsch, KL Kramer, E Ross, H Vongsachang, et al. 2015. The ontogeny of fairness in seven societies. *Nature* 528, 7581 (2015), 258–261.
[41] Alexander Bogner, Beate Littig, and Wolfgang Menz. 2009. *Interviewing experts*. Springer.
[42] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*. 4349–4357.
[43] Vani K Borooah, Amaresh Dubey, and Sriya Iyer. 2007. The effectiveness of jobs reservation: caste, religion and economic status in India. *Development and change* 38, 3 (2007), 423–445.
[44] C Boyes-Watson. 2014. Suffolk University, College of Arts & Sciences. *Center for Restorative Justice. Retrieved on November* 28 (2014), 2015.
[45] Eric Brewer, Michael Demmer, Bowei Du, Melissa Ho, Matthew Kam, Sergiu Nedevschi, Joyojeet Pal, Rabin Patra, Sonesh Surana, and Kevin Fall. 2005. The case for technology in developing regions. *Computer* 38, 6 (2005), 25–38.
[46] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. 77–91.
[47] Ministry of Statistics Central Statistics Office and Programme Implementation. 2018. *Women and Men in India: A statistical compilation of Gender related Indicators in India*. Technical Report. Government of India.
[48] Uma Chakravarti. 1993. Conceptualising Brahmanical patriarchy in early India: Gender, caste, class and state. *Economic and Political Weekly* (1993), 579–585.
[49] Maitrayee Chaudhuri. 2004. Feminism in India. (2004).
[50] 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).
[51] Andrew Cotter, Heinrich Jiang, Maya R Gupta, Serena Wang, Taman Narayan, Seungil 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.
[52] Kate Crawford. 2013. The hidden biases in big data. *Harvard business review* 1, 1 (2013), 814.
[53] Kate Crawford. 2013. Think again: Big data. *Foreign Policy* 9 (2013).
[54] William Crumpler. 2020. How Accurate are Facial Recognition Systems – and Why Does It Matter? | Center for Strategic and International Studies. (Accessed on 07/28/2020).
[55] Camera Culture. 2018. Economic Impact of Discoverability of Localities and Addresses in India — Emerging Worlds. http://mitemergingworlds.com/blog/2018/2/12/economic-impact-of-discoverability-of-localities-and-addresses-in-india. (Accessed on 09/24/2020).
[56] Abdi Lahir Dahir. 2019. Mobile loans apps Tala, Branch, Okash face scrutiny in Kenya — Quartz Africa. https://qz.com/africa/1712796/mobile-loans-apps-tala-branch-okash-face-scrutiny-in-kenya/. (Accessed on 08/04/2020).
[57] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516* (2019).
[58] The Living New Deal. [n.d.]. African Americans. https://livingnewdeal.org/what-was-the-new-deal/new-deal-inclusion/african-americans-2/. (Accessed on 08/29/2020).
[59] Mark Diaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing Age-Related Bias in Sentiment Analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173986
[60] Neha Dixit. July. Fair, But Not So Lovely: India's Obsession With Skin Whitening | by Neha Dixit | BRIGHT Magazine. https://brightthemag.com/fair-but-not-so-lovely-indias-obsession-with-skin-whitening-beauty-body-image-bleaching-4d6ba9c9743d. (Accessed on 09/25/2020).
[61] Pranav Dixit. 2019. India Is Creating A National Facial Recognition System. https://www.buzzfeednews.com/article/pranavdixit/india-is-creating-a-national-facial-recognition-system-and. (Accessed on 08/30/2020).

[62] Roel Dobbe, Sarah Dean, Thomas Gilbert, and Nitin Kohli. 2018. A broader view on bias in automated decision-making: Reflecting on epistemology and dynamics. *arXiv preprint arXiv:1807.00553* (2018).

[63] Jonathan Donner. 2015. *After access: Inclusion, development, and a more mobile Internet.* MIT press.

[64] Jonathan Donner, Nimmi Rangaswamy, M Steenson, and Carolyn Wei. 2008. "Express yourself " / "Stay together": Tensions surrounding mobile communication in the middle-class Indian family. *J. Katz (Ed.), Handbook of mobile communication studies* (2008), 325–337.

[65] Kevin P Donovan. 2015. The biometric imaginary: Bureaucratic technopolitics in post-apartheid welfare. *Journal of Southern African Studies* 41, 4 (2015), 815–833.

[66] Ariel Dorfman and Armand Mattelart. 1975. *How to Read Donald Duck.* International General New York.

[67] Susan Dray, Ann Light, A Dearden, Vanessa Evers, Melissa Densmore, D Ramachandran, M Kam, G Marsden, N Sambasivan, T Smyth, et al. 2012. Human–Computer Interaction for Development: Changing Human–Computer Interaction to Change the World. In *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications, Third Edition.* CRC press, 1369–1394.

[68] Esther Duflo. 2005. Why political reservations? *Journal of the European Economic Association* 3, 2-3 (2005), 668–678.

[69] Bert D'espallier, Isabelle Guérin, and Roy Mersland. 2011. Women and repayment in microfinance: A global analysis. *World development* 39, 5 (2011), 758–772.

[70] Arturo Escobar. 2011. *Encountering development: The making and unmaking of the Third World.* Vol. 1. Princeton University Press.

[71] Indian Express. 2020. Most Indian Nobel winners Brahmins: Gujarat Speaker Rajendra Trivedi. https://indianexpress.com/article/cities/ahmedabad/most-indian-nobel-winners-brahmins-gujarat-speaker-rajendra-trivedi-6198741/. (Accessed on 09/04/2020).

[72] Frantz Fanon. 2007. *The wretched of the earth.* Grove/Atlantic, Inc.

[73] Thomas B Fitzpatrick. 1988. The validity and practicality of sun-reactive skin types I through VI. *Archives of dermatology* 124, 6 (1988), 869–871.

[74] Rikin Gandhi, Rajesh Veeraraghavan, Kentaro Toyama, and Vanaja Ramprasad. 2007. Digital green: Participatory video for agricultural extension. In *2007 International conference on information and communication technologies and development.* IEEE, 1–10.

[75] Harris Gardiner. 2013. 5 in New Delhi Rape Case Face Murder Charges - The New York Times. https://www.nytimes.com/2013/01/04/world/asia/murder-charges-filed-against-5-men-in-india-gang-rape.html. (Accessed on 09/13/2020).

[76] Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.* 219–226.

[77] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010* (2018).

[78] Michael Golebiewski and Danah Boyd. 2019. Data voids: Where missing data can easily be exploited. *Data & Society* (2019).

[79] Masahiro Goto, Fuhito Kojima, Ryoji Kurata, Akihisa Tamura, and Makoto Yokoo. 2017. Designing matching mechanisms under general distributional constraints. *American Economic Journal: Microeconomics* 9, 2 (2017), 226–62.

[80] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology.* Vol. 47. Elsevier, 55–130.

[81] Ben Green. 2020. The false promise of risk assessments: epistemic reform and the limits of fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* 594–606.

[82] Ben Green and Salomé Viljoen. 2020. Algorithmic realism: expanding the boundaries of algorithmic thought. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* 19–31.

[83] Akhil Gupta. 2012. *Red tape: Bureaucracy, structural violence, and poverty in India.* Duke University Press.

[84] Alexa Hagerty and Igor Rubinov. 2019. Global AI Ethics: A Review of the Social Impacts and Ethical Implications of Artificial Intelligence. *arXiv* (2019), arXiv–1907.

[85] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* 501–512.

[86] Kurtis Heimerl, Shaddi Hasan, Kashif Ali, Eric Brewer, and Tapan Parikh. 2013. Local, sustainable, small-scale cellular networks. In *Proceedings of the Sixth International Conference on Information and Communication Technologies and Development: Full Papers-Volume 1.* 2–12.

[87] Virginia Held et al. 2006. *The ethics of care: Personal, political, and global.* Oxford University Press on Demand.

[88] David A Hollinger. 1998. *Science, Jews, and secular culture: studies in mid-twentieth-century American intellectual history.* Princeton University Press.

[89] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* 1–16.

[90] Ben Hutchinson and Margaret Mitchell. 2019. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency.* 49–58.

[91] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities. *ACL* (2020).

[92] IDSN. 2010. Two thirds of India's Dalits are poor - International Dalit Solidarity Network. https://idsn.org/two-thirds-of-indias-dalits-are-poor/. (Accessed on 08/13/2020).

[93] IEEE. 2019. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. "Classical Ethics in A/IS". In *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition.* 36–67.

[94] S Inayatullah. 2006. Culture and Fairness: The Idea of Civilization Fairness. In *Fairness, Globalization and Public Institutions.* University of Hawaii Press, 31–33.

[95] Azra Ismail and Neha Kumar. 2018. Engaging solidarity in data collection practices for community health. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–24.

[96] Mayank Jain. 2016. India's internet population is exploding but women are not logging in. *Scroll.in* (26 9 2016). https://scroll.in/article/816892/indias-internet-population-is-exploding-but-women-are-not-logging-inia

[97] Rob Jenkins and Anne Marie Goetz. 1999. Accounts and accountability: theoretical implications of the right-to-information movement in India. *Third world quarterly* 20, 3 (1999), 603–622.

[98] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.

[99] Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2016. Rawlsian fairness for machine learning. *arXiv preprint arXiv:1610.09559* 1, 2 (2016).

[100] Divij Joshi. 2020. AI Observatory. http://ai-observatory.in/. (Accessed on 12/30/2020).

[101] Yuvraj Joshi. 2018. Racial Indirection. *UCDL Rev.* 52 (2018), 2495.

[102] Rishi Ranjan Kala. 2019. High gender disparity among internet users in India - The Financial Express. https://www.financialexpress.com/industry/high-gender-disparity-among-internet-users-in-india/1718951/. (Accessed on 10/06/2020).

[103] Nathan Kallus and Angela Zhou. 2018. Residual unfairness in fair machine learning from prejudiced data. *arXiv preprint arXiv:1806.02887* (2018).

[104] Shivaram Kalyanakrishnan, Rahul Alex Panicker, Sarayu Natarajan, and Shreya Rao. 2018. Opportunities and Challenges for Artificial Intelligence in India. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society.* 164–170.

[105] Yuichiro Kamada and Fuhito Kojima. 2015. Efficient matching under distributional constraints: Theory and applications. *American Economic Review* 105, 1 (2015), 67–99.

[106] Anant Kamath and Vinay Kumar. 2017. In India, Accessible Phones Lead to Inaccessible Opportunities. https://thewire.in/caste/india-accessible-phones-still-lead-inaccessible-opportunities. (Accessed on 01/14/2021).

[107] Divya Kandukuri. 2018. Casteist Slurs You Need To Know - YouTube. https://www.youtube.com/watch?v=wJwkIxOpqZA. (Accessed on 09/25/2020).

[108] Kavita Karan. 2008. Obsessions with fair skin: Color discourses in Indian advertising. *Advertising & society review* 9, 2 (2008).

[109] Michael Katell, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Bintz, Daniella Raz, and PM Krafft. 2020. Toward situated interventions for algorithmic equity: lessons from the field. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* 45–55.

[110] Patrick Gage Kelley, Yongwei Yang, Courtney Heldreth, Christopher Moessner, Aaron Sedley, Andreas Kramm, David Newman, and Allison Woodruff. 2019. "Happy and Assured that life will be easy 10years from now.": Perceptions of Artificial Intelligence in 8 Countries. *arXiv preprint arXiv:2001.00081* (2019).

[111] Rachna Khaira. 2020. Surveillance Slavery: Swachh Bharat Tags Sanitation Workers To Live-Track Their Every Move | HuffPost India. https://www.huffingtonpost.in/entry/swacch-bharat-tags-sanitation-workers-to-live-track-their-every-move_in_5e4c98a9c5b6b0f6bff11f9b?guccounter=1. (Accessed on 07/28/2020).

[112] Srinivas Kodali. 2020. Aarogya Setu: A bridge too far? | Deccan Herald. https://www.deccanherald.com/specials/sunday-spotlight/aarogya-setu-a-bridge-too-far-835691.html. (Accessed on 08/01/2020).

[113] Ava Kofman. 2016. How Facial Recognition Can Ruin Your Life – Intercept. https://theintercept.com/2016/10/13/how-a-facial-recognition-mismatch-can-ruin-your-life/. (Accessed on 07/30/2020).

[114] Nitin Kohli, Renata Barreto, and Joshua A Kroll. 2018. Translation tutorial: a shared lexicon for research and practice in human-centered software systems. In *1st Conference on Fairness, Accountability, and Transparency. New York, NY, USA,* Vol. 7.

[115] Neha Kumar and Richard J Anderson. 2015. Mobile phones for maternal health in rural India. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems.* 427–436.

[116] Sunil Mitra Kumar. 2013. Does access to formal agricultural credit depend on caste? *World Development* 43 (2013), 315–328.

[117] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2016. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236* (2016).

[118] Michael Kwet. 2019. Digital colonialism: US empire and the new imperialism in the Global South. *Race & Class* 60, 4 (2019), 3–26.

[119] Jonas Lerman. 2013. Big data and its exclusions. *Stan. L. Rev. Online* 66 (2013), 55.

[120] Kwok Leung and Walter G Stephan. 2001. Social Justice from a Cultural Perspective. (2001).

[121] Kristian Lum and William Isaac. 2016. To predict and serve? *Significance* 13, 5 (2016), 14–19.

[122] Donald J Lund, Lisa K Scheer, and Irina V Kozlenkova. 2013. Culture's impact on the importance of fairness in interorganizational relationships. *Journal of International Marketing* 21, 4 (2013), 21–43.

[123] Ruth Macklin. 2004. *Double standards in medical research in developing countries*. Vol. 2. Cambridge University Press.

[124] Subramaniam Madheswaran and Paul Attewell. 2007. Caste discrimination in the Indian urban labour market: Evidence from the National Sample Survey. *Economic and political Weekly* (2007), 4146–4153.

[125] Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 615–621.

[126] Vidushi Marda. 2018. Artificial intelligence policy in India: a framework for engaging the limits of data-driven decision-making. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 2133 (2018), 20180087.

[127] Vidushi Marda and Shivangi Narayan. 2020. Data in New Delhi's predictive policing system. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 317–324.

[128] Donald Martin Jr, Vinod Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S Isaac. 2020. Participatory Problem Formulation for Fairer Machine Learning Through Community Based System Dynamics. *ICLR Workshop on Machine Learning in Real Life (ML-IRL)* (2020).

[129] Emma. Martinho-Truswell, Hannah. Miller, Isak Nti Asare, Andre Petheram, Richard (Oxford Insights) Stirling, Constanza Gómez Mont, and Cristina (C Minds) Martinez. 2018. Towards an AI strategy in Mexico: Harnessing the AI revolution. In *AI whitepaper*.

[130] Rachel Masika and Savita Bailur. 2015. Negotiating women's agency through ICTs: A comparative study of Uganda and India. *Gender, Technology and Development* 19, 1 (2015), 43–69.

[131] Achille Mbembe. 2015. Decolonizing knowledge and the question of the archive.

[132] Indrani Medhi, Aman Sagar, and Kentaro Toyama. 2006. Text-free user interfaces for illiterate and semi-literate users. In *2006 international conference on information and communication technologies and development*. IEEE, 72–82.

[133] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).

[134] Walter Mignolo. 2011. *The darker side of western modernity: Global futures, decolonial options*. Duke University Press.

[135] Government of India Ministry of Home Affairs. [n.d.]. 2011 Census Data. https://www.censusindia.gov.in/2011-Common/CensusData2011.html. (Accessed on 08/26/2020).

[136] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[137] Shakir Mohamed, Marie-Therese Png, and William Isaac. 2020. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology* (2020), 1–26.

[138] Angel Mohan. 2018. Why Urban Indian Women Turn Down Job Opportunities Away From Home |. https://www.indiaspend.com/why-urban-indian-women-turn-down-job-opportunities-away-from-home-94002/. (Accessed on 09/25/2020).

[139] Chandra Talpade Mohanty. 2005. *Feminism without borders: Decolonizing theory, practicing solidarity*. Zubaan.

[140] Anahita Mukherji. [n.d.]. The Cisco Case Could Expose Rampant Prejudice Against Dalits in Silicon Valley. https://thewire.in/caste/cisco-caste-discrimination-silicon-valley-dalit-prejudice. (Accessed on 08/14/2020).

[141] Geoff Mulgan and Vincent Straub. [n.d.]. The new ecosystem of trust: how data trusts, collaboratives and coops can help govern data for the maximum public benefit | Nesta. https://www.nesta.org.uk/blog/new-ecosystem-trust/. (Accessed on 08/21/2020).

[142] Deirdre K Mulligan, Joshua A Kroll, Nitin Kohli, and Richmond Y Wong. 2019. This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–36.

[143] Anand Murali. 2019. How India's data labellers are powering the global AI race | FactorDaily. https://factordaily.com/indian-data-labellers-powering-the-global-ai-race/. (Accessed on 09/13/2020).

[144] Lisa P Nathan, Michelle Kaczmarek, Maggie Castor, Shannon Cheng, and Raquel Mann. 2017. Good for Whom? Unsettling Research Practice. In *Proceedings of the 8th International Conference on Communities and Technologies*. 290–297.

[145] Newslaundry and Oxfam India. 2019. Who Tells Our Stories Matters: Representation of Marginalised Caste Groups in Indian Newsrooms. (8 2019).

[146] Helen Nissenbaum. 1996. Accountability in a computerized society. *Science and engineering ethics* 2, 1 (1996), 25–42.

[147] Rodrigo Ochigame. 2020. The Long History of Algorithmic Fairness. *Phenomenal World* (2020).

[148] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2 (2019), 13.

[149] Joyojeet Pal. 2008. Computers and the promise of development: aspiration, neoliberalism and "technolity" in India's ICTD enterprise. *A paper presented at confronting the Challenge of Technology for Development: Experiences from the BRICS* (2008), 29–30.

[150] Joyojeet Pal. 2015. Banalities turned viral: Narendra Modi and the political tweet. *Television & New Media* 16, 4 (2015), 378–387.

[151] Lawrence A Palinkas, Sarah M Horwitz, Carla A Green, Jennifer P Wisdom, Naihua Duan, and Kimberly Hoagwood. 2015. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and policy in mental health and mental health services research* 42, 5 (2015), 533–544.

[152] Rohini Pande. 2003. Can mandated political representation increase policy influence for disadvantaged minorities? Theory and evidence from India. *American Economic Review* 93, 4 (2003), 1132–1151.

[153] Kundan Pandey. 2020. COVID-19 lockdown highlights India's great digital divide. https://www.downtoearth.org.in/news/governance/covid-19-lockdown-highlights-india-s-great-digital-divide-72514. (Accessed on 01/14/2021).

[154] Priti Patnaik. 2012. Social audits in India – a slow but sure way to fight corruption. https://www.theguardian.com/global-development/poverty-matters/2012/jan/13/india-social-audits-fight-corruption. (Accessed on 08/21/2020).

[155] Amy Paul, C Jolley, and Aubra Anthony. 2018. Reflecting the Past, Shaping the Future: Making AI Work for International Development. *USAID. gov* (2018).

[156] Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation Sensitivity Analysis to Detect Unintended Model Biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5744–5749.

[157] Ashwin Rajadesingan, Ramaswami Mahalingam, and David Jurgens. 2019. Smart, Responsible, and Upper Caste Only: Measuring Caste Attitudes through Large-Scale Analysis of Matrimonial Profiles. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 393–404.

[158] Usha Ramanathan. 2014. Biometrics use for social protection programmes in India–Risk: Violating human rights of the poor. *United Nations Research Institute for Social Development* 2 (2014).

[159] Usha Ramanathan. 2015. Considering Social Implications of Biometric Registration: A Database Intended for Every Citizen in India [Commentary]. *IEEE Technology and Society Magazine* 34, 1 (2015), 10–16.

[160] C. Rangarajan, S. Mahendra Dev, K. Sundaram, Mahesh Vyas, and K.L Datta. 2014. *Report of the Expert Group to Review the Methodology for Measurement of Poverty*. Technical Report. Government of India Planning Commission.

[161] Rebecca Ratclifee. 2019. How a glitch in India's biometric welfare system can be lethal | India | The Guardian. https://www.theguardian.com/technology/2019/oct/16/glitch-india-biometric-welfare-system-starvation. (Accessed on 07/29/2020).

[162] Sharmila Rege. 1998. Dalit women talk differently: A critique of 'difference' and towards a Dalit feminist standpoint position. *Economic and Political Weekly* (1998), WS39–WS46.

[163] World Bank Human Development Unit South Asia Region. 2009. People with Disabilities in India: From Commitments to Outcomes. http://documents1.worldbank.org/curated/en/577801468259486686/pdf/502090WP0Peopl1Box0342042B01PUBLIC1.pdf. (Accessed on 08/26/2020).

[164] S. Henry Richardson. 2012. Fairness and Political Equality: India and the U.S. https://law.utah.edu/event/fairness-and-political-equality-india-and-the-u-s/.

[165] Sarah T Roberts. 2016. Digital refuse: Canadian garbage, commercial content moderation and the global circulation of social media's waste. *Wi: journal of mobile media* (2016).

[166] Valerian Rodrigues. 2011. Justice as the Lens: Interrogating Rawls through Sen and Ambedkar. *Indian Journal of Human Development* 5, 1 (2011), 153–174.

[167] Heather M Roff. 2020. Expected utilitarianism. *arXiv preprint arXiv:2008.07321* (2020).

[168] Oliver Rowntree. 2020. The mobile gender gap report 2020.

[169] Arundhati Roy. 2014. *Capitalism: A ghost story*. Haymarket Books.

[170] RT. 2017. 'Whoever leads in AI will rule the world': Putin to Russian children on Knowledge Day — RT World News. https://www.rt.com/news/401731-ai-rule-world-putin/. (Accessed on 09/20/2020).

[171] Cynthia Rudin and Joanna Radin. 2019. Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Science Review* 1, 2 (2019).

[172] Anouk Ruhaak. [n.d.]. Mozilla Foundation - When One Affects Many: The Case For Collective Consent. https://foundation.mozilla.org/en/blog/when-one-affects-many-case-collective-consent/. (Accessed on 08/21/2020).

[173] Rukmini S. 2020. India's poor are also document-poor. https://www.livemint.com/news/india/india-s-poor-are-also-document-poor-11578300732736.html. (Accessed on 09/13/2020).

[174] Rukmini S. May. In India, who speaks in English, and where? https://www.livemint.com/news/india/in-india-who-speaks-in-english-and-where-1557814101428.html. (Accessed on 09/25/2020).

[175] Nithya Sambasivan. 2019. The remarkable illusions of technology for social good. *interactions* 26, 3 (2019), 64–66.

[176] Nithya Sambasivan and Paul M Aoki. 2017. Imagined Connectivities: Synthesized Conceptions of Public Wi-Fi in Urban India. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 5917–5928.

[177] Nithya Sambasivan, Amna Batool, Nova Ahmed, Tara Matthews, Kurt Thomas, Laura Sanely Gaytán-Lugo, David Nemer, Elie Bursztein, Elizabeth Churchill, and Sunny Consolvo. 2019. " They Don't Leave Us Alone Anywhere We Go" Gender and Digital Abuse in South Asia. In *proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[178] Nithya Sambasivan, Garen Checkley, Amna Batool, Nova Ahmed, David Nemer, Laura Sanely Gaytán-Lugo, Tara Matthews, Sunny Consolvo, and Elizabeth Churchill. 2018. " Privacy is not for me, it's for those rich women": Performative Privacy Practices on Mobile Phones by Women in South Asia. In *Fourteenth Symposium on Usable Privacy and Security ({SOUPS} 2018)*. 127–142.

[179] Nithya Sambasivan, Ed Cutrell, Kentaro Toyama, and Bonnie Nardi. 2010. Intermediated technology use in developing communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2583–2592.

[180] Nithya Sambasivan and Jess Holbrook. 2018. Toward responsible AI for the next billion users. *interactions* 26, 1 (2018), 68–71.

[181] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

[182] Nithya Sambasivan and Thomas Smyth. 2010. The human infrastructure of ICTD. In *Proceedings of the 4th ACM/IEEE international conference on information and communication technologies and development*. 1–9.

[183] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1668–1678.

[184] Nadia Saracini and Murali Shanmugavelan. 2019. BOND: Caste and Development. (2019).

[185] Marie Schäfer, Daniel BM Haun, and Michael Tomasello. 2015. Fair is not fair everywhere. *Psychological science* 26, 8 (2015), 1252–1260.

[186] Amartya Kumar Sen. 2009. *The idea of justice*. Harvard University Press.

[187] Sungyong Seo, Hau Chan, P Jeffrey Brantingham, Jorja Leap, Phebe Vayanos, Milind Tambe, and Yan Liu. 2018. Partially generative neural networks for gang crime classification with partial information. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 257–263.

[188] Aabid shafi. 2018. Disability rights: Wheelchair users cannot access most of Delhi's buses. https://scroll.in/roving/894005/in-photos-why-wheelchair-users-in-delhi-find-it-difficult-to-use-buses-even-low-floor-ones. (Accessed on 09/25/2020).

[189] Shreya Shah. [n.d.]. #MissionCashless: Few use mobiles, fewer know what internet is in adivasi belts of Madhya Pradesh. https://scroll.in/article/824882/missioncashless-few-use-mobiles-fewer-know-what-internet-is-in-adivasi-belts-of-madhya-pradesh. (Accessed on 08/14/2020).

[190] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536* (2017).

[191] Murali Shanmugavelan. 2018. Everyday Communicative Practices of Arunthathiyars: The Contribution of Communication Studies to the Analysis of Caste Exclusion and Subordination of a Dalit Community in Tamil Nadu, India. (2018).

[192] Donghee (Don) Shin. 2019. Toward Fair, Accountable, and Transparent Algorithms: Case Studies on Algorithm Initiatives in Korea and China. *Javnost - The Public* 26, 3 (2019), 274–290. https://doi.org/10.1080/13183222.2019.1589249 arXiv:https://doi.org/10.1080/13183222.2019.1589249

[193] Ranjit Singh. 2018. 'The Living Dead'. *Whispers from the Field: Ethnographic Poetry and Creative Prose* (2018), 29–31.

[194] Ranjit Singh and Steven J Jackson. 2017. From Margins to Seams: Imbrication, Inclusion, and Torque in the Aadhaar Identification Project. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 4776–4824.

[195] Dean Spade. 2015. *Normal life: Administrative violence, critical trans politics, and the limits of law*. Duke University Press.

[196] Ajantha Subramanian. 2015. Making merit: The Indian Institutes of Technology and the social life of caste. *Comparative Studies in Society and History* 57, 2 (2015), 291.

[197] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976* (2019).

[198] Ranjula Bali Swain and Fan Yang Wallentin. 2009. Does microfinance empower women? Evidence from self-help groups in India. *International review of applied economics* 23, 5 (2009), 541–556.

[199] Nisha Tamang. 2020. Section 377: Challenges and Changing Perspectives in the Indian Society. *Changing Trends in Human Thoughts and Perspectives: Science, Humanities and Culture Part I* (2020), 68.

[200] Divy Thakkar, Nithya Sambasivan, Purva Kulkarni, Pratap Kalenahalli Sudarshan, and Kentaro Toyama. 2018. The Unexpected Entry and Exodus of Women in Computing and HCI in India. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[201] David R Thomas. 2006. A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation* 27, 2 (2006), 237–246.

[202] Sukhadeo Thorat and Paul Attewell. 2007. The legacy of social exclusion: A correspondence study of job discrimination in India. *Economic and political weekly* (2007), 4141–4145.

[203] Deeptiman Tiwary. 2015. Almost 68 percent inmates undertrials, 70 per cent of convicts illiterate | The Indian Express. https://indianexpress.com/article/india/india-news-india/almost-68-inmates-undertrials-70-of-convicts-illiterate/. (Accessed on 07/28/2020).

[204] Kentaro Toyama. 2015. *Geek heresy: Rescuing social change from the cult of technology*. PublicAffairs.

[205] Tunisia. 2018. National AI Strategy: Unlocking Tunisia's capabilities potential. http://www.anpr.tn/national-ai-strategy-unlocking-tunisias-capabilities-potential/. In *AI workshop*.

[206] Mazar Ullah. [n.d.]. Court told design flaws led to Bhopal leak | Environment | The Guardian. https://www.theguardian.com/world/2000/jan/12/1. (Accessed on 08/21/2020).

[207] Carol Upadhya. 2007. Employment, exclusion and'merit'in the Indian IT industry. *Economic and Political weekly* (2007), 1863–1868.

[208] Rajesh Veeraraghavan. 2013. Dealing with the digital panopticon: the use and subversion of ICT in an Indian bureaucracy. In *Proceedings of the Sixth International Conference on Information and Communication Technologies and Development: Full Papers-Volume 1*. 248–255.

[209] Srinivasan Vivek, Narayanan Rajendran, Chakraborty Dipanjan, Veeraraghavan Rajesh, and Vardhan Vibhore. 2018. Are technology-enabled cash transfers really 'direct'? *Economic and Political Weekly* 53, 30 (2018).

[210] Ngugi Wa Thiong'o. 1992. *Decolonising the mind: The politics of language in African literature*. East African Publishers.

[211] Immanuel Wallerstein. 1991. World system versus world-systems: A critique. *Critique of Anthropology* 11, 2 (1991), 189–194.

[212] Yilun Wang and Michal Kosinski. 2018. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of personality and social psychology* 114, 2 (2018), 246.

[213] Jayapal website. 2020. Jayapal Joins Colleagues In Introducing Bicameral Legislation to Ban Government Use of Facial Recognition, Other Biometric Technology - Congresswoman Pramila Jayapal. https://jayapal.house.gov/2020/06/25/jayapal-joins-rep-pressley-and-senators-markey-and-merkley-to-introduce-legislation-to-ban-government-use-of-facial-recognition-other-biometric-technology/. (Accessed on 07/30/2020).

[214] Maranke Wieringa. 2020. What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 1–18.

[215] Virginius Xaxa. 2011. Tribes and social exclusion. *CSSSC-UNICEF Social Inclusion Cell, An Occasional Paper* 2 (2011), 1–18.

[216] Alice Xiang and Inioluwa Deborah Raji. 2019. On the Legal Compatibility of Fairness Definitions. *arXiv preprint arXiv:1912.00761* (2019).

[217] Bendert Zevenbergen. 2020. *Internet Users as Vulnerable and at-Risk Human Subjects: Reviewing Research Ethics Law for Technical Internet Research*. Ph.D. Dissertation. University of Oxford. Unpublished PhD thesis.

[218] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 335–340.

[219] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457* (2017).

[220] Ran Zmigrod, Sebastian J Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1651–1661.