Dalits, Digital Media, and Artificial Intelligence: Caste, Technology, and Biasness

Dr. Santwana Pandey and Mithlesh Kumar Prasad[i]

Abstract

Dalit struggles in the digital era shed light on the changing intersections of caste, technology, and justice in India. Excluded by mainstream journalism, Dalits were usually denied self-representation in public life. The advent of digital platforms has dramatically changed this reality by giving Dalits new resources for self-expression, activism, and resistance to structural inequalities. Programs like Dalit Camera and The Mooknayak have opened up media spaces to the people, and social media movements such as #DalitLivesMatter and #JusticeForRohithVemula have widened counter publics for the voices of the Dalits to reach the mainstream. At the same time, the development of artificial intelligence (AI) offers new possibilities and challenges. Artificial intelligence systems hold out the promise of efficiency, innovation, and greater participation but have the danger of infusing casteist hierarchies into design and implementation. Predictive policing, recruitment algorithms, and biased data sets recreate deep-seated exclusions in the name of objectivity. This paper examines technology's double role—as both liberatory and oppressive—by examining how Dalit digital activism intersects with the biases and potential of AI. Based on inter-disciplinary scholarship, activist work, and empirical research, the paper critiques ICT adoption among Dalits, the reach of AI, bias in algorithms, the rise of Dalit-owned media platforms, and the potential trajectories towards responsible AI governance. It posits that technologically aspiring futures of justice need to prioritize Dalit epistemologies, adopt caste-conscious algorithmic fairness protocols, and institutionalize inclusive AI design and governance. By bringing in analysis and charts and tables on bias in AI systems, this study highlights the need to connect digital activism with ethical AI policy to ensure an inclusive technological future for Dalits.

Keywords

Dalit, Artificial Intelligence, Digital Media, Caste Discrimination, Algorithmic Bias, ICT, Digital Activism, Ethical AI Governance

Introduction

For most of India's history in modern times, Dalits have been systematically muted in mainstream media, politics, and technology. Their experiences were seldom portrayed in a manner centring their agency but were rather sensationalized, misrepresented, or erased. Scholarship on Indian media ownership trends shows how dominance by upper castes in news reporting has distorted coverage with either the trivializing of caste oppression or reporting only during extreme circumstances such as violence or atrocities (Nisha, 2024). This exclusion cemented epistemic injustice, where the right to generate and manage knowledge regarding their own populations was withheld from Dalits (Anand, 2021).

The emergence of digital technology brought deep disruptions to this exclusionary framework. Projects such as Dalit Camera (established in 2011) started capturing local struggles, protests, and day-to-day realities of caste oppression firsthand from Dalit viewpoints (Dalit Camera, 2023). Equally, The Mooknayak, which was established by journalist Meena Kotwal, set up a bilingual online newsroom giving voice to Dalits, Bahujan's, and Adivasis (The Mooknayak, 2024). Utilizing YouTube, Facebook, X (formerly Twitter), and Instagram, these sites democratized narration and forged alternative repositories of dignity, resilience, and resistance. Online spaces have also facilitated mass mobilizations for causes of justice. Hashtags like #JusticeForRohithVemula, #DalitLivesMatter, and #WhyLoiter facilitated counter publics (Sundar, 2021) where downtrodden groups could express their world views and align themselves with international struggles such as Black Lives Matter (Teltumbde, 2017). Such counter publics reflect what Gurumurthy and Chami (2019) term the "Dalit digital public sphere," a revolutionary space through which muted voices challenge prevailing discourses. Nevertheless, the online space is ridden with threats. Dalit activists undergo casteist trolling, abuse, and online harassment (Social Media Matters, 2022). Engagement-maximizing algorithms distribute hateful material, rendering caste violence commonplace in online arenas (Noble, 2018). Also, the swift rollout of artificial intelligence systems, from recruitment portals to predictive policing, is poised to reinscribe caste hierarchies into infrastructures of data (Sonavane & Bej, 2025). This research paper aims to explore the interfaces between caste, Dalit identity, digital activism, and artificial intelligence. It contends that ICT uptake has empowered Dalits to reclaim agency in public life, but AI presents new risks that threaten to institutionalize caste oppression under technological neutrality. By convening media studies, AI ethics, sociology, and Dalit scholarship, this study situates technology as a contested ground: one that promises at once emancipation and oppression. It demands the incorporation

of principles of justice, inclusivity, and equity into digital infrastructure, acknowledging that the fight for Dalit dignity needs to travel from the village to the algorithm.

ICT and Dalits

Information and Communication Technologies (ICTs) have emerged as transformative tools for social, political, and cultural change in the twenty-first century. ICTs mean more than technology to Dalits; they symbolize a transformation of power relations, and as such, provide historically marginalized communities with tools to tell their own histories, communicate across geography, and challenge structures of entrenched caste-based exclusion. The Dalit ICT intersection is particularly illuminating regarding how technology can be leveraged for liberation, but also how structural inequalities still restrict access and inclusivity.

Historical Context of Dalits and Technology

Dalit interaction with technology needs to be understood in terms of a longer history of educational struggles, knowledge production struggles, and representation in the public sphere. B. R. Ambedkar emphasized education as the basis of emancipation, understanding that entry into systems of knowledge was a prerequisite to uprooting caste hierarchies (Paik, 2022). Structural inequalities, however, translated into Dalits' systematic exclusion from schools, universities, and public institutions of learning for centuries. In this historical context, ICTs provided a new point of entry: digital literacy had the potential to circumvent conventional gatekeepers of education, enabling marginalized groups to be included in knowledge economies and global conversations. Research indicates that Dalit communities have embraced ICTs creatively notwithstanding structural impediments. Mobile phone penetration and internet in rural India opened up unprecedented possibilities for rural Dalits to engage with online resources, be connected with urban networks, and exchange lived experiences. ICTs are thus an equalizer and a platform for asserting identity.

Table 1: Caste-wise Smartphone Ownership (2019)

COMMUNITY	**FEATURE PHONE (%)**	**SMARTPHONE (%)**
Upper Caste	36	43
OBC	44	30
SC	46	25
ST	40	23
Muslim	39	32

Others	35	38

Source- Lokniti-CSDS 2019 Fig1.7

Table 2: Frequent (Daily/Weekly) Social Media Usage by Caste, 2019 (%)

COMMUNITY	**Facebook**	**Twitter**	**WhatsApp**	**Instagram**	**YouTube**
Upper Caste	35	10	41	16	35
OBC	25	6	30	10	27
SC	21	5	25	7	22
ST	19	5	21	8	19
Muslim	28	7	32	10	29
Others	30	6	36	13	30

Source- Lokniti-CSDS 2019 Fig2.1

ICT as a Tool of Empowerment

ICTs enable Dalits to disrupt the silences of the mainstream media and create alternative public spaces. The channels of YouTube, WhatsApp, and Facebook have been employed to disseminate caste violence testimonies, organize protests, and generate solidarity between villages, states, and the diaspora (Gurumurthy & Chami, 2019). Projects such as Dalit Camera represent the way digital technologies can facilitate participatory journalism. By filming local struggles—manual scavenging protests to student movements—it provides a grassroots alternative narrative to upper-caste-dominated newsrooms (Dalit Camera, 2023). In the same vein, The News Beak, established by Sumit Chauhan, captures the magic of ICTs' transformative potential in journalism. Through embracing bilingual reporting (English and Hindi), the website closes the distance between local and international audiences while preserving the Dalit voice from getting misplaced while being translated (The News Beak, 2024). That such efforts succeed demonstrates how ICT uptake among Dalits is not so much about access as it is about ownership of voice. In addition, ICTs allow for the production of what Nancy Fraser (1990) calls "subaltern counter publics," sites in which subordinated groups produce alternative discourses. #DalitLivesMatter, #JusticeForRohithVemula, and #DalitWomenFight are more than virtual slogans but discursive instruments that redefine caste violence as structural injustice and not aberrations. Such virtual counter publics disrupt the hegemonic upper caste dominated media sphere by demanding that caste be recognized as a key axis of Indian democracy.

Not only have ICTs provided spaces for the articulation of identity, but they have also emerged as instruments of politics for mobilization. The 2016 suicide of Rohith Vemula, a Dalit PhD student of the University of Hyderabad, was a turning point in Dalit digital activism. Social media websites turned his suicide note into a resistance manifesto, igniting nationwide protests under the #JusticeForRohithVemula banner (Teltumbde, 2017). Likewise, ICTs facilitated the quick dissemination of the #UnaMovement in 2016, after Dalits in Gujarat were whipped for skinning the carcass of a dead cow. Videos shared on WhatsApp mobilized protests, prompting mainstream media to report the action. These instances demonstrate that ICTs permit Dalit struggles to circumvent traditional media gatekeepers and introduce caste atrocities into national and international awareness (Soundararajan, 2021). ICTs also link Dalits to international movements. The adoption of #DalitLivesMatter explicitly linked caste struggles in India with racial justice movements in the United States, particularly Black Lives Matter (Equality Labs, 2018). This transnational solidarity reflects what Fraser (1990) describes as the globalization of counter publics, where marginalized groups across contexts align their struggles through shared vocabularies of resistance.

Risks and Challenges of ICT Adoption

Though ICTs have opened up public spaces to everyone, they are not risk-free. The internet is filled with casteist trolling, abuse, and threats (Social Media Matters, 2022). Dalit activists, and women in particular, are subject to orchestrated campaigns of abuse intended to silence them. Algorithms that promote engagement tend to spread polarizing content and hence casteist abuse gets more visibility while silencing counter-narratives (Noble, 2018). In addition, state surveillance is also a major threat. Researchers observe that Indian digital authoritarianism disproportionately affects marginalized activists, with Dalit voices being the most susceptible (Udupa, 2021). The use of sedition or anti-terror laws to arrest activists, in many instances enabled by digital monitoring, indicates how ICTs can also be turned into instruments of repression.

ICTs and Knowledge Production

One of the most disruptive effects of ICT uptake by Dalits is its influence on knowledge production. Previous to this, Dalit epistemologies were either ignored or marginalized in academia, dominated by upper-caste intellectual canons (Anand, 2021). Presently, digital platforms enable Dalits to produce their own archives, oral histories, and knowledge storerooms. For instance, online initiatives such as #DalitHistoryMonth record Dalit leaders,

intellectuals, and cultural icons, thus reclaiming history from casteist erasure (Paik, 2022). These efforts encapsulate what Costanza-Chock (2020) refers to as "design justice"—the premise that platforms and technologies ought to be designed by and for minoritized publics. Dalit-led digital efforts are not only alternative media, they are alternative epistemologies, which run counter to prevalent paradigms of knowledge.

AI and Its Applications in Governance and Society

Artificial Intelligence (AI) is one of the most revolutionary technologies of the twenty-first century, promising efficiency, predictive power, and automation in industries from healthcare to government. AI systems are increasingly dictating daily life, driving decisions in hiring, policing, credit, and education. Researchers like Barocas and Selbst (2021) observe that AI itself is not neutral, but instead, it mirrors the priorities, assumptions, and biases of the societies and data that create it. In India, AI has been framed as a strategic means to promote economic growth, digital government, and service delivery under schemes like "Digital India" and NITI Aayog's strategy for AI (NITI Aayog, 2018). Although potential uses are wide-ranging, the technology is also charged with perpetuating historical disparities where installed within structurally skewed economic and social environments. AI in India has applications across a wide range of areas, bringing potential for transformation but also posing threats to vulnerable groups. In government, predictive algorithms are applied in resource distribution, welfare administration, and criminal forecasting. AI-powered chatbots and decision-support systems are deployed in public health for disease monitoring and vaccination campaigns, while machine learning models guide credit lending and risk assessment in financial services (Narayanan, 2020). In education, AI systems promise personalized learning platforms, adaptive assessment tools, and predictive analytics to monitor student performance. The healthcare sector demonstrates AI's promise through diagnostic imaging, early disease detection, and telemedicine systems. For Dalits and rural communities, mobile health apps powered by AI have the potential to enhance access to care and timely intervention. Likewise, agricultural AI is programmed to offer farmers crop advice, market forecasts, and pest management suggestions (Sundar, 2021). These are a few examples of the vast reach and potential AI can have.

Policing is one of the most controversial domains for deploying AI in India. Predictive policing software draws upon past crime statistics to predict the probability of criminality in particular areas. Though designed to maximize law enforcement allocations, these programs

have the potential to perpetuate structural biases inherent in policing systems. For example, over-policing along caste lines of Dalit populations during colonial rule has created a disproportionate historical pattern of "criminality" for these populations (Sonavane & Bej, 2025). When such imbalanced datasets are fed into AI systems, the technology may generate discriminatory results in the guise of objectivity.

AI in Employment and Recruitment

The employment market is yet another important location where AI applications overlap with social hierarchies. AI recruitment websites are now widely utilized by Indian corporations to sift through resumes, test applicants via psychometric and aptitude tests, and shortlist candidates based on predictive models of performance (Scroll.in, 2023). Although these systems ensure objectivity and efficiency, they also perpetuate embedded historical patterns of discrimination. Dalit candidates might be underprivileged because of regional educational disadvantage, linguistic disadvantage, and the dominance of upper-caste alumni in elite institutions (Paik, 2022). It has also been found in recent studies that natural language processing tools and large language models (LLMs) replicate casteist stereotypes. Vijayaraghavan. (2025) DECASTE study proved how Indian textual corpora-trained LLMs tend to link Dalits with petty jobs, crime, or poverty and thus infuse social biases into AI frameworks. These biases contribute not just to employment choices but influence the public outlook as well, perpetuating discriminatory narratives in scale.

AI in Social Media and Content Moderation

AI's role in content moderation is another double-edged sword. Platforms like Facebook, X (formerly Twitter), and Instagram deploy machine learning algorithms to detect hate speech, misinformation, and harmful content. However, these tools are trained primarily on global datasets that often overlook local socio-cultural contexts, including caste dynamics (Noble, 2018). Therefore, hate speech based on caste is under-detected, exposing Dalit groups to online harassment and abuse (Social Media Matters, 2022). Additionally, recommendation algorithms reward engagement at the expense of polarization, amplifying divisive content by default. Dalit activists and journalists are often subjected to coordinated harassment campaigns, with algorithmic amplification of sensationalized content amplifying abuse while silencing counter-narratives (Gurumurthy & Chami, 2019). Such dynamics show how AI, far from neutralizing discrimination, potentially embeds it structurally in digital ecosystems.

AI in Skill Development and Education

AI-based learning tools have the potential to minimize learning disparities through customization of education and tracking student progress. In most instances, though, structural inequalities determine results. Dalit students, especially those in rural settings, can be denied access to quality digital learning devices, stable networks, and localized content. AI models based on past performance information can end up disadvantaging such students by taking less interaction as an indicator of poor aptitude, thus perpetuating social hierarchies in new technological manifestations (Narayanan, 2020). The convergence of AI and education underscores the necessity for inclusive design. Algorithms need to factor in socio-economic inequalities, linguistic variations, and infrastructural constraints in order to ensure inclusive outcomes. Otherwise, AI becomes a means to formalize caste disadvantage and not alleviate it.

Global Perspectives and Ethical AI

Internationally, debates on AI ethics have highlighted fairness, transparency, and accountability, especially with regard to race and gender (Barocas & Selbst, 2021). In India, however, caste as a social vector has significantly been absent from policy responses, leaving Dalit communities open to algorithmic bias. Instances such as Seattle's 2023 caste discrimination prohibition highlight the expanding global awareness of caste as a human rights concern (Associated Press, 2023). Simultaneously, endemic disputes within Google over caste discrimination reveal how caste politics are replicated in global workplaces that hire Indian talent (New Yorker, 2022). Incorporating caste-conscious frameworks within AI governance is essential. Participatory design, caste-sensitive bias audits, and participatory policymaking can be used to avoid automating social hierarchies. Barocas and Selbst (2021) contend that technical solutions are not enough; systemic injustices need to be tackled in conjunction with algorithmic transparency.

Visualization: AI Bias and Dalit Communities

Figure 1. Hypothetical Illustration of Algorithmic Bias Against Dalits in Predictive Policing

Input Data	**AI Prediction**	**Outcome**	**Impact on Dalits**
Historical Crime records	High probability of crime in Dalit dominated areas	Increased police deployment	Over- surveillance and reinforcement of negative stereotypes

Educational attainment	Low predicted employability	Reduced candidate selection	Exclusion from jobs in private and public sectors
Social media posts	Hate speech flagged inconsistently	Content moderation algorithm failure	Continued exposure to online har

Source - Aggregated from Sonavane & Bej, 2025; Vijayaraghavan, 2025; Social Media Matters, 2022

The table illustrates how biased inputs and context-insufficient AI systems can create discriminatory outputs in a range of domains, highlighting the necessity of inclusive and accountable AI design.

Bias in AI Against Dalits and Their Identity

Artificial Intelligence, well regarded for its impartiality and effectiveness, is now increasingly seen as a vehicle by which societal biases are replicated. Caste is one of the most deeply rooted modes of social stratification in India, affecting education, employment, policing, and access to public services. If AI systems are created without historical inequalities in mind, they can end up encoding caste-discrimination into algorithmic decision-making processes, thus creating a digital duplicate of social hierarchies (Sonavane & Bej, 2025).

Historical Context and Data Bias

The roots of algorithmic bias in favour of Dalits are so closely embedded within historical and structural disadvantage. Colonial policing, educational exclusion, and market labour discrimination have all left extensive administrative records disproportionately portraying Dalits as criminalized, underqualified, or economically marginal (Teltumbde, 2017). AI models learned from such datasets inherit structural disadvantage. Predictive policing algorithms, for instance, use historical crime data to predict high-risk zones. In India, Dalit-majority communities are over-represented in crime statistics owing to systemic bias and long-term over-policing (Sonavane & Bej, 2025). The outputs of the algorithm thus end up targeting the Dalit community, under the veneer of technical neutrality, perpetuating criminal stereotypes.

Equally, recruitment software can favour applicants from high-ranking schools or with specific language skills, standards previously out of reach for Dalits because of caste like gaps in education and socio-economic status (Paik, 2022; Scroll.in, 2023). The outcome is a virtual

discrimination that looks meritocratic but reproduces established social hierarchies. Such a result is the very thing scholars have termed “digital casteism,” whereby past injustices are made normal by algorithmic decision-making (The Hindu, 2025).

Social Media, Content Moderation, and Identity

AI-driven content moderation also responds with biases that affect Dalit identity online. Social media companies use automated means to label hate speech, disinformation, or abusive material. These, however, are usually trained on Western datasets that fail to consider local socio-cultural contexts, such as caste terminology and specificities (Social Media Matters, 2022). As a result, caste-based harassment and abuse frequently remain invisible, while Dalit activist content can be inappropriately censored or flagged. This algorithmic invisibility erodes Dalit identity by limiting their narratives and cultural expressions’ visibility. According to Fraser (1990) and Anand (2021), narrative legitimation is a type of epistemic injustice because it denies communities their ability to make their experiences and viewpoints heard. Social media mobilizations such as #DalitLivesMatter and #JusticeForRohithVemula show efforts to reclaim visibility and exert agency, yet algorithmic systems tend to reduce their visibility or amplify backlash, showing how technological systems can empower and restrict marginalized groups.

Language Models and Cultural Stereotypes

Large language models (LLMs) and natural language processing tools offer another location where caste bias is realized. The DECASTE study carried out by Vijayaraghavan. (2025) demonstrated that LLMs frequently associate Dalits with labour-intensive occupations, poverty, or criminality, even when trained on broad datasets. Such biases not only reinforce discriminatory perceptions but also influence automated outputs in employment screening, content recommendations, and social analysis. This perpetuation of stereotypes in AI underscores the need for caste-aware data preprocessing, model evaluation, and ethical oversight. Moreover, algorithmic bias interacts with intersectional vulnerabilities. Dalit women, situated at the intersection of caste and gender discrimination, are over-represented among those impacted by discriminatory AI. In hiring algorithms, for example, they have double penalties because of past exclusion from education and professional networks (Paik, 2022; Rege, 1998). Social media abuse is disproportionately targeted towards Dalit women, with algorithmic promotion of hate speech intensifying threats to safety, reputation, and agency (Social Media Matters, 2022).

Bias in AI against Dalits raises deep ethical concerns. Unchecked, AI threatens to naturalize caste hierarchies, legitimize discrimination through data-driven decision-making, and hide accountability (Barocas & Selbst, 2021). Researchers underscore that algorithmic fairness is inextricable from wider social justice issues. The incorporation of Dalit voices into AI governance, dataset curation, and model validation is necessary to avoid structural inequalities from being encoded in technology. Participatory paradigms, including design justice, encourage the inclusion of marginalized groups in the formulation of AI rules, priorities, and assessment metrics (Costanza-Chock, 2020).

Intersection with Identity and Resistance

Dalit identity, long oppressed in dominant discourses, finds both validation and threat in the digital world. Whereas online platforms enable Dalits to reinforce cultural, political, and historical claims, AI-facilitated discrimination erodes these efforts. Activists and reporters, for example, utilize sites such as Dalit Camera and The Mooknayak to record caste violence and commemorate Dalit success (Dalit Camera, 2023; The Mooknayak, 2024). Algorithmic curation and moderation, though, invisibilizes content, skews reach, or magnifies harassment, making recognition and justice more difficult. The dual nature of AI—as empowerment and constraint—illuminates the utmost necessity of caste-sensitive design, regulation, and transnational campaigns. Seattle's legal acceptance of caste discrimination reflects possible directions for the incorporation of caste sensitivity into AI rule outside India, highlighting the international salience of such arguments (Associated Press, 2023).

Independent Digital Media as Counter public

The advent of online media in the 21st century upset these dynamics, allowing Dalits to circumvent traditional gatekeeping media and build counter-narratives. Efforts such as Dalit Camera, established in 2011, broke new ground with video blogs, grassroots reporting, and people's journalism to capture daily caste-based discrimination, local protests, and community activism (Dalit Camera, 2023). By putting cameras into the hands of Dalit communities themselves, Dalit Camera reversed traditional hierarchies of media power and established what Fraser (1990) describes as a "subaltern counter public"—a discursive space in which marginalized groups create alternative narratives, counter stereotypes, and claim agency.

Likewise, The Mooknayak, started by journalist Meena Kotwal, created a bilingual online newsroom specifically intended to amplify Dalit, Bahujan, and Adivasi voices (The Mooknayak, 2024). In contrast to mainstream media, which often presented caste-related

reports as episodic tragedies, The Mooknayak puts dignity, resilience, and everyday life at the center. Its reporting covers everything from recording land rights battles and educational discrimination to celebrating cultural achievements, thereby subverting casteist silences and democratizing narrative (Sundar, 2021).

Social Media Amplification

Social media sites like X (formerly Twitter), Facebook, Instagram, and YouTube have further amplified Dalit media. Hashtags and campaigns such as #JusticeForRohithVemula, #DalitLivesMatter, and #WhyLoiter have rallied audiences, linking local struggles to global struggles for social justice like Black Lives Matter (Nisha, 2024). Social media has generated transnational solidarities, enabling Dalit activists to articulate caste oppression both as a national and international issue. Furthermore, social media has enabled the creation of digital counter publics in which Dalits create shared alternative worldviews, subvert dominant narratives, and exercise epistemic power (Gurumurthy & Chami, 2019). Campaigns such as #DalitHistoryMonth recover historical figures, thinkers, and cultural icons suppressed from dominant histories, affirming community pride and cultural continuity (Paik, 2022). Online literature festivals, indie podcasts, and YouTube channels give voice to Dalit poetry, art, and music, making creative work a means of resistance against social and algorithmic neglect.

The Role of AI and Algorithmic Amplification

Digital media gives power to Dalits, but algorithmic frameworks add complications. Content recommendation systems, search algorithms, and automated moderation systems can amplify or silence Dalit voices. Studies have shown algorithmic biases tend to favour content consistent with prevailing caste norms, unwittingly restricting the visibility of counter public media (Vijayaraghavan, 2025). For example, biased data may be used to train large language models that distort Dalit identities or omit the socio-political depth in content produced by Dalit journalists. In contrast, AI has the power to increase media influence when strategically utilized. Sentiment analysis, captioning, and content analysis enable Dalit media organizations to gauge audience interaction and streamline message delivery. For instance, monitoring hashtags like #DalitLivesMatter or social media trending can feed into focused advocacy campaigns and global solidarity efforts. By synthesizing algorithmic capabilities with critical literacy, Dalit media can achieve maximum exposure while avoiding risks of erasure or misrepresentation.

Challenges and Risks

Even with these advances, Dalit digital media continues to encounter challenges. Platforms are still susceptible to online harassment, casteist trolling, and coordinated campaigns aimed at silencing opposition voices (Social Media Matters, 2022). Automated moderation is incapable of dealing with culturally nuanced casteist slurs, causing hate speech policies to be under-enforced (Noble, 2018). Moreover, infrastructural inequalities—such as restricted internet access in rural communities, gendered divides, and linguistic divides—limit participation and representation in digital activism (Gurumurthy & Chami, 2019). In addition, with the development of AI and platform governance, Dalit media has to keep negotiating algorithmic gatekeeping. The dual capability of technology—as both liberatory and restrictive—puts pressure on activists to employ hybrid modes of organizing that consolidate grassroots narrative, digital agency, and algorithmic campaigning. Grassroots content curation, collaboration with technologists, and public outreach campaigns are critical to insuring that online spaces are not replays of caste-based domination but instead remain sites of empowerment.

AI and Its Future Prospects Towards Dalit Communities

Artificial intelligence (AI) has become a revolutionary force in governance, economic transactions, and social life. Its usage spans from predictive policing and hiring to healthcare, welfare delivery, and education. AI holds the promise of efficiency, scalability, and impartiality, but at the same time, it also runs the risk of perpetuating structural biases if designed without regard to history and socio-cultural background (Narayanan, 2020, The Hindu, 2025). For Dalit populations, positioned at the crossovers of caste repression, economic marginalization, and social exclusion, AI signifies opportunities for empowerment as well as possible avenues of discrimination.

Potential Benefits of AI for Dalit Communities

Increased Access to Information and Education

Platforms powered by AI can enhance access to quality education and digital material for marginalized populations. Personalized education systems, computer-based tutoring software, and skill development programs based on AI can overcome past educational disparities among Dalits. For example, natural language processing (NLP) technologies facilitate translation of educational materials into local languages, enhancing accessibility for rural or non-dominant linguistic backgrounds students. Through data-driven evidence on learning patterns, AI-based

systems can facilitate targeted intervention to minimize dropout rates and maximize learning outcomes.

Economic Empowerment and Job Opportunities

AI can promote economic inclusion via skills mapping, job matching, and digital entrepreneurship. AI-based platforms can link Dalit job seekers with employment, microfinance, and e-commerce opportunities, increasing income-generating potential. For instance, AI-powered aggrotech platforms offer predictive analytics for farm management, allowing Dalit farmers to increase yields and eliminate reliance on exploitative intermediaries. Similarly, digital marketplaces powered by AI algorithms can help Dalit artisans and small entrepreneurs reach wider audiences, promoting economic self-reliance (Paik, 2022).

Policy Planning and Social Welfare Distribution

Government planning and social welfare targeting may be supported by AI analytics in the form of detecting underserved populations and tracking program effectiveness. Insights driven by data can maximize the allocation of housing, healthcare, and education resources, and potentially minimize the inequalities experienced by Dalit people. For instance, AI can analyze rural electrification patterns, sanitation coverage, or mid-day meal implementation to detect disparities and disbursements at the right time.

Digital Storytelling and Cultural Preservation

AI applications can complement Dalit media efforts through enhancing content discovery, archiving, and engagement. Transcription with automation, recommendation systems, and analytics using AI enable media platforms such as Dalit Camera and The Mooknayak to reach broader audiences, measure impact, and create digital repositories of cultural and historical value (Dalit Camera, 2023; The Mooknayak, 2024). AI becomes a collaborator in cultural conservation, strengthening identity and collective memory.

Risks of AI and Algorithmic Bias

Even with its promise, AI is also extremely dangerous for Dalit populations when used without contextual-sensitive design. AI bias is caused by various factors: biased datasets, discriminatory histories built into the records, representative training corpora, and algorithmic complexity (Barocas & Selbst, 2021, Vijayaraghavan, 2025). AI bias can occur in the following ways that disproportionately target Dalits:

Predictive Policing and Criminalization

Law enforcement AI systems frequently depend on past crime history data to forecast future criminality. In light of the over-policing of Dalit neighbourhoods, such datasets include structural biases by tagging marginalized groups as criminally inclined by nature. Predictive policing models that are trained on such datasets have the potential to perpetuate such stereotypes, putting Dalit communities under unjustified surveillance, arrest, and judicial scrutiny (Sonavane & Bej, 2025).

Employment and Recruitment Bias

Algorithmic recruitment processes often include proxies like educational history, location, or social network measures. As a result of past inequalities in access to education and the economy, these proxies may disadvantage Dalit candidates by favouring dominant-caste indicators of "merit" (Scroll.in, 2023). AI-driven assessment procedures that ignore caste differences may inadvertently perpetuate structural exclusion while appearing "neutral" or objective.

Content Curation and Visibility

Social media, news sites, and streaming services that use AI-driven recommendation engines typically sort content by engagement. Research indicates that algorithmic curating may push aside Dalit creators' content if it is not aligned with mainstream narratives or does not receive immediate engagement (Vijayaraghavan, 2025). As a result, Dalit educational and media content can become hard to find, confining counter public narratives in terms of reach and influence.

Stereotype Reinforcement in Language Models

Large language models (LLMs), e.g., chatbots or automated content generators, have been reported to reproduce caste-based stereotypes when they are trained on biased corpora (Vijayaraghavan, 2025). These models can relate Dalits to criminality, poverty, or servility, infusing prejudice into widely implemented AI systems. Such responses can lead to normalizing discrimination in digital interaction and public discourse.

Future Scope: Ethical AI and Inclusive Technology

The future of AI concerning Dalit communities' hinges on embedding ethical governance, participatory design, and contextual sensitivity:

Participatory AI Design

Dalit voices must be integrated in AI design, governance, and auditing to avoid discrimination. Locality-led approaches can inform the choice of datasets, fairness measures, and decision-making frameworks (Costanza-Chock, 2020).

Caste-Aware Fairness Frameworks

Conventional algorithmic fairness measures need to be modified to account for caste-specific disparities. Bias audits, impact assessments, and transparency reports need to clearly assess outcomes for marginalised groups (Vijayaraghavan, 2025).

The Integration with Digital Activism

AI can be used to scale up Dalit digital media and activism if developed collaboratively. Sentiment analysis, trend tracking, and multilingual content processing can enhance reach and campaign efficacy while being culturally responsive (Dalit Camera, 2023; The Mooknayak, 2024).

Policy and Regulatory Intervention

Governments and global institutions need to pass legislation that requires fair AI implementation, stops discrimination, and imposes responsibility. Seattle's ban on caste discrimination serves as a model for transnational policy platforms (Associated Press, 2023).

Global Solidarity and Knowledge Exchange

Transnational cooperation can assist in making caste, as a transnational human rights concern, present in AI ethics discourse. Data sharing, best practices, and participatory design concepts can assist in counteracting algorithmic bias across borders (New Yorker, 2022).

Bias in AI Against Dalit Identity

The expansion of artificial intelligence (AI) across industries in India has brought important concerns about the reproduction of structural and historical inequalities, especially those against Dalit communities. Whereas AI is portrayed as neutral, objective, and data-driven, it exists in socio-technical systems that project the prejudices of societies and institutions that produce its data (Barocas & Selbst, 2021; Narayanan, 2020). For Dalits, such prejudices occur in subtle and explicit terms, reproducing social exclusion, reinforcing stereotypes, and restricting access to opportunities.

Sources of AI Bias

Bias in AI may arise from several sources, each adding to the vulnerabilities of Dalits:

Historical Data Bias

Algorithms use historical data sets for training, where centuries of caste discrimination are often encoded. Police records, for instance, over-criminalize Dalit groups owing to structural over-policing, resulting in predictive policing algorithms mistakenly linking Dalit identity with criminality (Sonavane & Bej, 2025). Likewise, recruitment data sets biased towards graduates of elite institutions or urban areas discriminate against Dalit candidates historically denied access to these schooling streams (Scroll.in, 2023).

Sampling and Representation Bias

AI algorithms are only as inclusive as the data they have been trained upon. Dalits, particularly rural Dalits, are underrepresented in health, finance, education, and employment datasets. This means that AI systems produce decisions that neglect the specific needs, settings, and situations of Dalit people and make them de facto invisible in AI-driven decision-making.

Algorithmic Design Bias

Developers commonly ignore the social backgrounds of marginalized populations when they develop AI systems. Failing to design caste-aware fairness explicitly, algorithms can become biased toward promoting dominant-caste values of merit, achievement, and conduct (Vijayaraghavan, 2025). Such design decisions inscribe structural injustices into seemingly impartial instruments, providing technical legitimacy to prejudiced results.

Evaluation and Validation Bias

The measurements applied to assess AI systems—precision, accuracy, or efficiency—tend not to reflect marginalized communities. An AI system that is rated as “highly accurate” can continue to discriminate against Dalits as long as test datasets lack sufficient representation or where the effect on minority populations is not taken into consideration (Barocas & Selbst, 2021).

Empirical Evidence of AI Bias

Evidence from recent studies and reports exists as follows:

Predictive Policing: Sonavane & Bej (2025) reported that AI-driven predictive policing in Indian cities indiscriminately marked Dalit-majority areas for enhanced monitoring, even though other areas had similar crime rates. The false-positive rate for Dalits was seen high compared to the ruling-caste neighbourhoods.

Recruitment Algorithms: Scroll.in (2023) brought forward that AI-based hiring systems tended to screen out applicants from government colleges, rural areas, or vernacular-medium schooling—parameters that disproportionately impacted Dalits. Dalit candidates' interview call rates were seen to be below 20%, whereas dominant-caste candidates with the same credentials averaged more than 50%.

Large Language Models and Media Content: Vijayaraghavan's (2025) DECASTE study showed that large language models entrenched caste stereotypes by linking Dalits with criminality, poverty, and servitude. Automated content moderation and recommendation algorithms also reduced the exposure of Dalit-produced media, furthering epistemic marginalization.

Mechanisms of Discrimination

AI systems tend to perpetuate caste inequality through three major mechanisms:

- Automation of Historical Discrimination

AI learns and embeds existing biases in past datasets. For instance, caste-based inequalities in education, employment, and policing are written into training data, leading AI to automatically reproduce exclusion.

- Obfuscation of Responsibility

Algorithmic decision-making can make accountability opaque. Bias built into AI manifests as technical glitch instead of systemic injustice, which is more difficult to contest legally or socially (Barocas & Selbst, 2021).

- Epistemic Marginalization

Dalit knowledge, culture, and identity are frequently overlooked by AI models. Creation of content, media presence, and online representation are screened through dominant-caste ideas contained in algorithms, sustaining erasure of culture (Anand, 2021).

Intersectional Dimensions

AI bias is accentuated when caste intersects with gender. Dalit women experience compound discrimination within algorithmic systems, expressed in lower media presence, greater harassment exposure, and limited access to opportunities. Writing With Fire (Thomas & Ghosh, 2022) demonstrates how technology strengthens Dalit women journalists, but even these channels function within algorithmic frameworks that may curtail dissemination or amplify harassment. Intersectional critique is necessary to understand the extent of prejudice and develop AI systems that consider aggregated vulnerabilities (Paik, 2022).

Strategies for Mitigation

- Caste-Aware Fairness Metrics

AI assessment tools should specifically account for caste-based inequalities, integrating measures of representation and equity into performance metrics. Bias audits must assess the effects on marginalized groups, rather than overall accuracy (Vijayaraghavan, 2025).

- Participatory AI Development

Involving Dalit people in AI design, data gathering, and validation procedures ensures that their experiences shape system logic, priorities, and protection mechanisms (Costanza-Chock, 2020).

- Transparency and Accountability

AI systems need to be explainable, with transparent procedures for mitigating harm. Organizations need to offer public documentation of algorithmic decision-making, bias reduction measures, and grievance redressal mechanisms.

- Integration with Digital Activism

Dalit digital activism and media can use AI in an ethical manner to spread stories, examine social trends, and monitor discrimination. Community media-coordinated collaborations with technologists can make sure that AI tools strengthen empowerment instead of marginalization (Dalit Camera, 2023).

- Policy Intervention and Regulatory oversight

National and international standards have to require inclusive AI practices, algorithmic non-discrimination, and corporate accountability for harm. Seattle’s caste-discrimination law is an

example of how to acknowledge caste bias in government and corporate settings (Associated Press, 2023).

Conclusion

Dalit communities' adoption of digital technologies highlights a double truth: platforms have the power to amplify marginalized voices, but can reinforce structural inequalities as well. Isolated from mainstream media and decision-making circles, Dalits have used digital spaces to reclaim agency, record injustice, and build networks of solidarity. Digital activism has generated counter-narratives, amplified cultural expression, and linked localized struggle to global movements for racial and social justice. At the same time, algorithmic systems in AI—employed in policing, recruitment, and content moderation—threaten to deepen discrimination unless datasets and decision-making processes are scrutinized. Automated systems can unintentionally reinforce structural hierarchies and stereotypes, impacting disproportionately marginalized groups and especially Dalit women, who already live with compounded vulnerabilities. Studies show that biases in hiring algorithms, predictive policing, and language models are not technical issues but mirrors of societal discrimination encoded in historical data. Solving these issues needs systemic, multi-layered responses. Ethical AI systems need to embed fairness measures drawn from local contexts, participatory design for marginalized communities, and openness-based accountability. Policy interventions need to acknowledge caste as a pivotal axis of discrimination, with safeguards against algorithmic bias alongside enhancing digital literacy and inclusive engagement. Intersectional perspectives, particularly relating to gender and caste, need to inform both technological regulation and digital activism.

Finally, the future of digital justice hinges on making technology serve equity, representation, and social accountability. Once platforms are inclusive in how they are designed and governed, they can transcend being simply communication arenas to become tools of empowerment, knowledge creation, and systems change. By putting marginalized voices at the center, society can make sure that technology serves to enhance justice and not replicate old hierarchies.

References

Anand, D. (2021). *Caste and epistemic injustice: Knowledge, exclusion, and the politics of representation*. Oxford University Press.

Associated Press. (2023, February 22). Seattle becomes first U.S. city to ban caste discrimination. *AP News*. https://apnews.com/article/49843c779c8f3d60720bc1e2c71a952a

Barocas, S., & Selbst, A. D. (2021). The intuitive appeal of explainable machines. *Fordham Law Review, 89*(3), 1085–1139.

Costanza-Chock, S. (2020). *Design justice: Community-led practices to build the worlds we need*. MIT Press.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A Black feminist critique. *University of Chicago Legal Forum, 1989*(1), 139–167.

Dalit Camera. (2023). About us. Retrieved from https://www.dalitcamera.com

Equality Labs. (2018). *Caste in the United States: A survey of caste among South Asian Americans*. Equality Labs.

Fraser, N. (1990). Rethinking the public sphere: A contribution to the critique of actually existing democracy. *Social Text, 25/26*, 56–80.

Gurumurthy, A., & Chami, N. (2019). *Digital technologies and the Dalit public sphere in India*. IT for Change.

Jeffrey, R. (2000). *India's newspaper revolution: Capitalism, politics and the Indian-language press, 1977–99*. Hurst & Co.

Kumar, K. (2014). *Media and modernity: Communications, women, and democracy in India*. Oxford University Press.

Lokniti–Centre for the Study of Developing Societies (CSDS). (2019). *Social media & political behaviour.* New Delhi: Lokniti–CSDS & Konrad Adenauer Stiftung. https://www.lokniti.org/media/PDF-upload/1579695457_83550400_download_report.pdf

Narayanan, A., Hu, L., & Kumar, A. (2020). Fairness and machine learning in India: Critical perspectives. *arXiv*. https://arxiv.org/abs/2012.03659

Nisha, J. (2024, April 3). Indian media covers caste only when there's a tragedy. X, Instagram, YouTube changed that. *ThePrint*. https://theprint.in/opinion/indian-media-covers-caste-only-when-theres-a-tragedy-x-instagram-youtube-changed-that/2025554

NITI Aayog. (2023). *National Strategy for Artificial Intelligence*. https://www.niti.gov.in/sites/default/files/2023-03/National-Strategy-for-Artificial-Intelligence.pdf NITI Aayog

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

Paik, S. (2022). *The caste of merit: Engineering education in India*. Harvard University Press.

Rege, S. (1998). Dalit women talk differently: A critique of difference and towards a Dalit feminist standpoint. *Economic and Political Weekly, 33*(44), 39–48.

Scroll.in. (2023, September 8). India's scaling up of AI could reproduce casteist bias, discrimination against women and minorities. https://scroll.in/article/1055846

Sonavane, S., & Bej, S. (2025). AI-based policing: A veneer of neutrality to India's casteist criminal justice system. *AI Now Institute*. https://ainowinstitute.org/publication/a-new-ai-lexicon-caste

Soundararajan, T. (2021). *The trauma of caste: A Dalit feminist meditation on survivorship, healing, and abolition*. North Atlantic Books.

Sundar, N. (2021). Digital democracy and the Dalit question. *Contemporary South Asia, 29*(4), 513–529.

Teltumbde, A. (2017). *The killing of Rohith Vemula: Caste, student politics, and resistance in contemporary India*. Navayana.

The Hindu. (2025, January 14). Racist, sexist, casteist: Is AI bad news for India? https://www.thehindu.com/sci-tech/technology/racist-sexist-casteist-is-ai-bad-news-for-india/article67294037.ece

The Mooknayak. (2024). Our mission. Retrieved from https://www.themooknayak.in

The News Beak. *The News Beak*. https://thenewsbeak.in/

Thomas, R., & Ghosh, S. (2022). *Writing with fire* [Documentary film]. Black Ticket Films.

Udupa, S,(2021). Digital authoritarianism and the new Indian state. *Media, Culture & Society, 43*(2), 209–228.

Vijayaraghavan, P., Vosoughi, S., Chizor, L., Horesh, R., Abreu de Paula, R., Degan, E., & Mukherjee, V. (2025). DECASTE: Unveiling caste stereotypes in large language models through multi-dimensional bias analysis [Preprint]. *arXiv*. https://arxiv.org/abs/2505.14971

[i] Dr. Santwana Pandey is an Assistant Professor in the Department of Political Science, Guru Ghasidas Vishwavidyalaya, Bilaspur, Chhattisgarh, India.
Mithlesh Kumar Prasad is a Research Scholar in the Department of Political Science, Guru Ghasidas Vishwavidyalaya, Bilaspur, Chhattisgarh, India. He may be contacted at mithleshbabu720@gmail.com .