VALUE ALIGNMENT IN THE GLOBAL SOUTH: A MULTIDIMENSIONAL APPROACH TO NORM ELICITATION IN INDIAN CONTEXTS

Atmadeep Ghoshal, Martim Brandão & Ruba Abu-Salma
Department of Informatics
King's College London
London, United Kingdom
{atmadeep.ghoshal,martim.brandao,ruba.abu-salma}@kcl.ac.uk

ABSTRACT

This paper addresses critical gaps in artificial intelligence (AI) value alignment research concerning historically marginalized communities in the Global South, with a specific focus on Dalits and Adivasis in India. We propose a multidimensional approach that integrates B.R. Ambedkar's and Amartya Sen's theoretical frameworks for social justice with Clifford Geertz's thick description methodology to develop context-sensitive norm elicitation processes. By examining how deeply entrenched sociopolitical hierarchies influence these communities' agency to process and express information about their own values, we demonstrate that conventional approaches to value alignment inadequately address the unique challenges faced by these communities. Our framework emphasizes the role of Indian AI missions in creating culturally relevant scenarios for norm elicitation, ensuring meaningful participation of marginalized communities in AI alignment processes. This approach not only advances the discourse on inclusive AI development, but also provides practical strategies for implementing value alignment methodologies that acknowledge and address historical power dynamics.

1 INTRODUCTION

Scholars have approached the problem of value alignment in AI systems from various directions, with recent sociotechnical investigations transforming the discourse through frameworks like STELA (Bergman et al., 2024), a framework focusing on aligning language models with human values in their broader social contexts, and Constitutional AI (Bai et al., 2022), which proposes training AI systems with explicit rules and constraints to ensure that they behave in accordance with specified values and principles. These frameworks have emphasized inclusivity and representational equity, addressing inherent biases and differential impacts on marginalized populations. However, current sociotechnical value alignment research has insufficiently addressed the complexities of historically marginalized communities in the Global South, particularly regarding epistemic injustice in norm elicitation methodologies. Epistemic injustice in norm elicitation occurs when certain communities' knowledge and perspectives are systematically excluded or undervalued in the process of determining what values should guide AI systems (Kay et al., 2024; Fricker, 2007). As Varshney observes, the conceptualization of marginality within Global South contexts demands nuanced analysis, necessitating the integration of decolonial frameworks into value alignment paradigms to address complex power dynamics (Varshney, 2023). This observation becomes particularly relevant when considering communities whose value systems have been shaped by historical oppression. As scholars like Spivak and Guha note, prolonged systemic marginalization can constrain the self-deterministic agency of at-risk groups, making their organic value orientations socially invisible (Guha, 1996; Morris, 2009). This presents a unique challenge for value alignment processes, particularly in the norm elicitation phase, where we attempt to understand and incorporate community values into AI systems. Our focus on Dalits and Adivasis in India provides a critical case study to examine these challenges with regard to norm elicitation. Drawing on social justice principles of BR Ambedkar (Ambedkar, 2016) and Amartya Sen (Sen, 2010), we propose that effective norm elicitation for value alignment must go beyond conventional methods. To this end, we advocate for the construction of scenarios that depict the impact of AI (adverse or otherwise) on these communities within the norm elicitation process as a helpful probe that can better capture their perspectives. This approach fundamentally

differs from existing norm elicitation methodologies in two critical dimensions: First, it shifts focus from primarily interaction-based alignments with large language models to a more comprehensive impact-based framework. Second, it calls for institutional mechanisms—such as national AI missions in India—to proactively uncover and articulate values that might otherwise remain obscured by historical power dynamics.

2 The Embedded Problematics of Decolonial AI

Recent scholarship has made progress in incorporating diverse perspectives into AI development through Human-Centered AI Interaction (HCAI) and Explainable and Responsible AI (XRAI) frameworks, with a particular focus on the Global South and India in domains such as education, healthcare, and public perceptions (Bingley et al., 2023; Koster et al., 2022; Reuel et al., 2024). The decolonial approach has emerged as a key framework for re-conceptualizing fairness and facilitating marginalized communities' participation in AI governance (Hao, 2020; Mhlambi & Tiribelli, 2023). However, existing approaches to norm elicitation overlook a critical challenge: the ability of at-risk groups to authentically articulate their values. This problem is particularly pronounced in the Indian context, where historical institutions like the caste system have systematically oppressed communities such as Dalits and Adivasis (Thorat, 2008). Postcolonial scholarship, drawing on the work of theorists like Spivak, reveals how subaltern groups' epistemological frameworks are mediated through hegemonic power structures (Morris, 2009). This epistemic dependency means that marginalized communities' worldviews and value systems are fundamentally shaped by historically dominant groups. The challenge extends beyond technological literacy or representation, questioning whether these communities can genuinely engage in norm elicitation processes that assume epistemic independence. The core issue is not only the explicit barriers to participation, but also the implicit cognitive frameworks that shape how these communities process and articulate their values (Davenport & Trivedi, 2013). Historical oppression has created a complex landscape in which the ability to express authentic preferences is compromised by deeply ingrained power dynamics (Pandey & Nagarkoti, 2021; Netto et al., 2021; Yeşilyurt & Vezne, 2023). To meaningfully include these communities in AI value alignment, we must first fully understand how historical oppression affects their ability to articulate their own values and norms. This requires a nuanced approach that acknowledges the intricate reality of epistemic dependency and seeks to surface authentic perspectives within this complex social context.

3 Indian Value Alignment: Foundations and Principles

Our proposed multidimensional framework for Indian value alignment is grounded in three key objectives. Firstly, it requires a philosophical foundation rooted in India's sociocultural, economic, and political contexts to define its moral trajectory. To achieve this, we draw on the seminal work of B.R. Ambedkar (Ambedkar, 2016) and Amartya Sen (Sen, 2010), whose extensive research on social justice and the empowerment of marginalized communities offers essential theoretical insights. Secondly, capturing the authentic perspectives of marginalized groups such as Dalits and Adivasis requires adopting Clifford Geertz's thick description methodology (Geertz, 1977). While primarily used for ethnographic fieldwork, thick description is invaluable for engaging with structural elements within communities, enabling a nuanced understanding of their decision-making processes and organic value systems (Luhrmann, 2015). This approach fosters a deeper engagement with their lived realities. Lastly, to ensure that these communities have the agency to participate in value alignment activities such as norm elicitation, we emphasize the need for Indian AI missions to adopt a proactive stance. This involves creating educational resources to inform subaltern populations about the potential adverse impacts of AI technologies if misaligned with their value systems and rigorously disseminating this knowledge. With these elements in place, norm-elicitation efforts in India would better reflect inclusivity and diversity.

3.1 Philosophical Foundations: Social Justice and Freedom

We would like to define the essence of morality in our framework through a social justice lens, particularly drawing on two seminal theoretical frameworks. This approach is necessitated by our understanding that marginalized communities require greater algorithmic representation in AI technologies, while previous research indicates that historically oppressed communities in the Global South need systematic empowerment to participate meaningfully in technological development processes. Our recommendations are anchored in the scholarship of B.R. Ambedkar's and Amartya

Sen. Ambedkar's theoretical framework for social justice, particularly his critique of caste-based hierarchies and emphasis on *annihilation of caste*, provides crucial insights into structural inequalities. His conception of social justice transcends mere political rights, emphasizing that true emancipation requires economic and social liberation. Ambedkar's work on institutional safeguards for Dalits and other marginalized communities not only has shaped India's constitutional principles but also offers valuable perspectives on representation and participatory democracy. Sen's theoretical contributions, particularly their critique of Rawlsian justice, offer complementary insights. While Rawls emphasizes the *veil of ignorance* and primary goods, Sen's capability approach argues that justice should be evaluated in terms of actual freedoms and opportunities available to individuals. His emphasis on *development as freedom* and critique of purely resource-based approaches to justice provide a theoretical foundation for understanding how technological interventions should enhance substantive freedoms rather than merely providing access. Ambedkar's and Sen's focus on agency and participation aligns with our framework's emphasis on enabling marginalized communities to be active participants in technological development rather than passive recipients.

3.2 Methodological Foundations: Thick Description

Recent research in AI safety and value alignment highlights that effectively eliciting norms necessitates meaningful engagement with a wide range of value systems, especially those of marginalized communities. Works by Huang et al. demonstrate how conventional approaches to AI alignment often fail to capture the nuanced value structures of marginalized populations, instead defaulting to surface-level preferences that can inadvertently reinforce existing power hierarchies (Hung, 2024). Similarly, Mohamed et al.'s research on decolonial AI emphasizes that robust alignment systems must engage with indigenous knowledge frameworks that have been historically suppressed in technological development (Mohamed et al., 2020). These insights become particularly crucial when considering communities like Dalits and Adivasis, whose relationship with technology could be argued to be mediated through layers of historical oppression following the subaltern dialogue (Vaghela et al., 2022) . As Sambasivan et al. observe in their analysis of India's AI ecosystem, traditional responsible AI approaches often fail to account for how caste hierarchies and structural inequalities influence the ability of these communities to articulate their values for AI technologies (Sambasivan et al., 2021). By extension, within the domain of value alignment, conventional approaches, whether through human feedback, preference learning, or reward modeling, can often be understood as operating under the assumption of autonomous agency, which may not hold for these communities whose very self-conception has been shaped by centuries of systemic oppression. In this regard, we propose the thick description methodology, developed by anthropologist Clifford Geertz, as our primary methodological framework for value alignment.

Although thick description has traditionally been used as an ethnographic tool, we argue that it offers unique advantages in uncovering and incorporating marginalized perspectives into AI alignment frameworks. Unlike conventional preference learning methods that prioritize explicit statements and observable behaviors, thick description offers a more nuanced epistemological approach. Where traditional methods may be employed to conduct structured interviews or surveys, thick description involves immersive ethnographic techniques: extended participant observation, in-depth narrative interviews, careful documentation of contextual interactions, and a rigorous interpretation of symbolic meanings. This means not just recording what people say, but meticulously analyzing how they say it, the broader cultural contexts that shape their statements, and the unspoken power dynamics that influence their expressions. By engaging with what Geertz terms the *multiplicity of complex conceptual structures*, thick description allows researchers to uncover layers of meaning that standard quantitative or surface-level qualitative methods may miss entirely. Thus, there are three key attributes of thick description that make it particularly suitable for value alignment objectives: First, its emphasis on understanding behaviors within their complete cultural context allows us to trace how historical oppression and power dynamics influence communities' conceptions of beneficial AI outcomes. Second, its focus on interpreting symbolic actions enables us to access indigenous knowledge systems that may contain crucial insights for alignment, but are not immediately apparent through direct questioning. Third, its recursive nature provides a methodological framework for understanding how individual experiences of marginalization should inform the development of AI safety measures. We believe that this approach would enrich the construction of scenarios as probes for norm elicitation by grounding them in the lived experiences of individuals from these communities. Our approach also stands in distinct contrast to traditional norm elicitation methods. Where traditional approaches may seek to directly elicit stated preferences for AI behavior, thick description allows us to understand how these preferences are embedded within larger systems of meaning and power. Where current methods may treat technological engagement as a straightforward process

of preference aggregation, thick description enables us to examine how historical marginalization shapes the very possibility of meaningful participation in alignment processes.

3.3 Operational Foundations: Proactive Role of Indian AI Missions

Drawing inspiration from Google's work on moral imagination for engineering teams (Keeling et al., 2024), we propose leveraging scenario-based probes to examine AI's impact on Dalit and Adivasi communities during the norm elicitation process. Our examination of existing literature and subaltern theory reveals that conventional approaches—standardized questionnaires or interaction-based safety evaluations—prove insufficient given these marginalized communities' constrained agency in self-expression. Such methodologies would not only narrow their focus to large language models (LLMs), but would also exclude other critical AI applications from consideration, including predictive policing, judicial decision-making systems, and professional candidate profiling. A major concern arises regarding the extent to which LLM value alignment knowledge can be effectively generalized to other technological domains, which could perpetuate harm to these communities. Therefore, we advocate for the development of rigorously researched scenarios, presented through infographics or audiovisual media, to illuminate both adverse and beneficial AI impacts. These scenario-based probes can serve as instruments of knowledge empowerment, enabling communities to comprehend technological implications—an approach substantiated by prior research in Human-Computer Interaction (HCI), particularly in studies involving robotics (Ashwini et al., 2024) and mobile health systems (Okolo et al., 2021). Within India's policy and development framework, several indigenous programs for AI and related technologies are designated as "missions." Notable among these are initiatives such as ***AI4Bharat*** [1] and the ***National Cyberphysical Systems Mission*** [2]. We posit that these missions should assume the critical responsibility of scenario construction through active engagement with marginalized communities to develop culturally nuanced scenarios that examine the implications of AI deployment across various domains. Key scenarios could explore how automated credit scoring systems can inadvertently encode caste markers through proxy variables; the potential perpetuation of discrimination through AI-enabled job screening tools drawing from historically biased data; the implications of AI-powered healthcare triage systems for communities with historically limited medical access; and the impact of automated content moderation on the digital representation of Dalit and Adivasi narratives.

The scenario construction process should follow a systematic methodology examining three key dimensions: the existing patterns of discrimination and exclusion; the technical architecture of proposed AI systems; and the points of intersection between these systems and marginalized communities' lived experiences. In educational contexts, for instance, scenarios may investigate how automated assessment systems interpret dialectical variations among first-generation learners, or how AI-driven personalized learning platforms may reinforce educational disparities due to limited digital resource access. These analyses can draw valuable insights from documented cases such as facial recognition systems' bias in Brazil (Peron & Evangelista, 2024) and the COMPAS recidivism algorithm's racial bias in the U.S. justice system (Engel et al., 2024). These carefully constructed scenarios could then subsequently inform norm elicitation processes and alignment activities, including reinforcement learning from human feedback in a democratic and inclusive manner.

4 Conclusion

Our investigation into value alignment challenges in the Global South reveals the complex interplay between historical oppression, epistemic dependency, and technological development. The proposed multidimensional framework, which combines social justice principles with thick description methodology and scenario-based assessments, offers a systematic approach to engaging with marginalized perspectives in AI development. By emphasizing the role of Indian AI missions in facilitating norm elicitation through culturally relevant scenarios, we provide a practical pathway for meaningful inclusion of Dalit and Adivasi voices in value alignment processes. This framework will enhance discussions on ethics while also laying the groundwork for creating genuinely inclusive AI systems that reflect and respect the values of historically marginalized communities. Future work should focus on implementing and evaluating the proposesd framework across different contexts, ensuring that AI development in the Global South genuinely serves the interests of its most at-risk populations.

[1] https://ai4bharat.iitm.ac.in/

[2] https://nmicps.in/

REFERENCES

Bhimrao Ramji Ambedkar. *Annihilation of caste*. Verso Books, London, England, January 2016.

B Ashwini, Atmadeep Ghoshal, Venkata Ratnadeep Suri, Krishnaveni Achary, and Jainendra Shukla. "It looks useful, works just fine, but will it replace me?" Understanding Special Educators' Perception of Social Robots for Autism Care in India. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642836. URL https://doi.org/10.1145/3613904.3642836.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI Feedback, 2022. URL https://arxiv.org/abs/2212.08073.

Stevie Bergman, Nahema Marchal, John Mellor, Shakir Mohamed, Iason Gabriel, and William Isaac. STELA: a community-centred approach to norm elicitation for AI alignment. *Scientific Reports*, 14(1), March 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-56648-4. URL http://dx.doi.org/10.1038/s41598-024-56648-4.

William J. Bingley, Caitlin Curtis, Steven Lockey, Alina Bialkowski, Nicole Gillespie, S. Alexander Haslam, Ryan K.L. Ko, Niklas Steffens, Janet Wiles, and Peter Worthy. Where is the human in human-centered AI? Insights from developer priorities and user experiences. *Computers in Human Behavior*, 141:107617, April 2023. ISSN 0747-5632. doi: 10.1016/j.chb.2022.107617. URL http://dx.doi.org/10.1016/j.chb.2022.107617.

Christian Davenport and Priyamvada Trivedi. Activism and awareness: Resistance, cognitive activation, and 'seeing' untouchability among 98,316 Dalits. *Journal of Peace Research*, 50 (3):369–383, May 2013. ISSN 1460-3578. doi: 10.1177/0022343313477885. URL http://dx.doi.org/10.1177/0022343313477885.

Christopher Engel, Lorenz Linhardt, and Marcel Schubert. Code is law: how COMPAS affects the way the judiciary handles the risk of recidivism. *Artificial Intelligence and Law*, April 2024. ISSN 1572-8382. doi: 10.1007/s10506-024-09400-2. URL http://dx.doi.org/10.1007/s10506-024-09400-2.

Miranda Fricker. *Epistemic Injustice*. Oxford University Press, June 2007. ISBN 9780198237907. doi: 10.1093/acprof:oso/9780198237907.001.0001. URL http://dx.doi.org/10.1093/acprof:oso/9780198237907.001.0001.

Clifford Geertz. *The interpretation of cultures*. Basic Books, London, England, April 1977.

Ranajit Guha. *Subaltern studies: Volume 1*. Subaltern Studies. OUP, Oxford, England, September 1996.

Karen Hao. The problems AI has today go back centuries. *MIT Technology Review.*, July 2020.

Kai-Hsin Hung. A case of assessing the working and living conditions of data workers in India's global artificial intelligence value chains. *SSRN Electronic Journal*, 2024. ISSN 1556-5068. doi: 10.2139/ssrn.4983127. URL http://dx.doi.org/10.2139/ssrn.4983127.

Jackie Kay, Atoosa Kasirzadeh, and Shakir Mohamed. Epistemic Injustice in Generative AI, 2024. URL https://arxiv.org/abs/2408.11441.

Geoff Keeling, Benjamin Lange, Amanda McCroskery, David Weinberger, Kyle Pedersen, and Ben Zevenbergen. Moral imagination for engineering teams: The technomoral scenario. *International Review of Information Ethics*, 34(1):1–8, 2024.

Raphael Koster, Jan Balaguer, Andrea Tacchetti, Ari Weinstein, Tina Zhu, Oliver Hauser, Duncan Williams, Lucy Campbell-Gillingham, Phoebe Thacker, Matthew Botvinick, and Christopher Summerfield. Human-centred mechanism design with Democratic AI. *Nature Human Behaviour*, 6(10):1398–1407, July 2022. ISSN 2397-3374. doi: 10.1038/s41562-022-01383-x. URL http://dx.doi.org/10.1038/s41562-022-01383-x.

Tanya M. Luhrmann. *Thick Description: Methodology*, pp. 291–293. Elsevier, 2015. ISBN 9780080970875. doi: 10.1016/b978-0-08-097086-8.44057-2. URL http://dx.doi.org/10.1016/B978-0-08-097086-8.44057-2.

Sábëlo Mhlambi and Simona Tiribelli. Decolonizing AI ethics: Relational autonomy as a means to counter AI harms. *Topoi*, 42(3):867–880, 2023. doi: 10.1007/s11245-022-09874-2.

Shakir Mohamed, Marie-Therese Png, and William Isaac. Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philos. Technol.*, 33(4):659–684, December 2020.

Rosalind Morris (ed.). *Can the subaltern speak?* Columbia University Press, New York, NY, September 2009.

Gina Netto, Lynne Baillie, Theodoros Georgiou, Lai Wan Teng, Noraida Endut, Katerina Strani, and Bernadette O'Rourke. Resilience, smartphone use and language among urban refugees in the Global south. *Journal of Ethnic and Migration Studies*, 48(3):542–559, June 2021. ISSN 1469-9451. doi: 10.1080/1369183x.2021.1941818. URL http://dx.doi.org/10.1080/1369183X.2021.1941818.

Chinasa T. Okolo, Srujana Kamath, Nicola Dell, and Aditya Vashistha. "It cannot do all of my work": Community Health Worker Perceptions of AI-Enabled Mobile Health Applications in Rural India. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445420. URL https://doi.org/10.1145/3411764.3445420.

Shubham Pandey and Priyanshi Nagarkoti. An anthropological analysis of the invisibility of Dalits of India in the environmental discourse: A tale of subjugation, alienation and resistance. *Contemporary Voice of Dalit*, 13(2):165–176, April 2021. ISSN 2456-0502. doi: 10.1177/2455328x211008366. URL http://dx.doi.org/10.1177/2455328X211008366.

Alcides Eduardo Dos Reis Peron and Rafael Evangelista. Beyond Instrumentarianism: Automated Facial Recognition Systems in Brazil and Digital Colonialism's Violence. *Science, Technology and Society*, 29(4):535–554, October 2024. ISSN 0973-0796. doi: 10.1177/09717218241281819. URL http://dx.doi.org/10.1177/09717218241281819.

Anka Reuel, Patrick Connolly, Kiana Jafari Meimandi, Shekhar Tewari, Jakub Wiatrak, Dikshita Venkatesh, and Mykel Kochenderfer. Responsible AI in the Global Context: Maturity Model and Survey, 2024. URL https://arxiv.org/abs/2410.09985.

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. Re-imagining Algorithmic Fairness in India and Beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 315–328, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445896. URL https://doi.org/10.1145/3442188.3445896.

Amartya Sen. *The Idea of Justice*. Penguin Books, Harlow, England, July 2010.

Sukhadeo Thorat. *Dalits in India*. SAGE Publications, Thousand Oaks, CA, December 2008.

Palashi Vaghela, Steven J Jackson, and Phoebe Sengers. Interrupting Merit, Subverting Legibility: Navigating Caste In 'Casteless' Worlds of Computing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3502059. URL https://doi.org/10.1145/3491102.3502059.

Kush R. Varshney. Decolonial AI Alignment: Openness, Visesa-Dharma, and Including Excluded Knowledges. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2023. URL https://api.semanticscholar.org/CorpusID:267069454.

Etem Yeşilyurt and Rabia Vezne. Digital literacy, technological literacy, and internet literacy as predictors of attitude toward applying computer-supported education. *Education and Information Technologies*, 28(8):9885–9911, January 2023. ISSN 1573-7608. doi: 10.1007/s10639-022-11311-1. URL http://dx.doi.org/10.1007/s10639-022-11311-1.