frontiers | Frontiers in Psychology

TYPE Opinion
PUBLISHED 03 November 2025
DOI 10.3389/fpsyg.2025.1657227

Whose bias gets coded? Psychology's role in decolonizing AI

Divya Lakshmi S[1]*, Visakh Mohan[2], P. S. Reeja[3], Elizabeth Alexander[4], Neethu Varghese[5] and Alphonsa Kurian[3]

[1]Department of Computer Applications, Marian College Kuttikkanam Autonomous, Idukki, India, [2]School of Social Work, Marian College Kuttikkanam Autonomous, Idukki, India, [3]Department of Sociology, Bishop Chulaparambil Memorial College, Kottayam, India, [4]PG Department of Social Work, KE College, Kottayam, India, [5]Department of English, Bishop Chulaparambil Memorial College, Kottayam, India

Check for updates

OPEN ACCESS

EDITED BY
Kostas Karpouzis,
Panteion University, Greece

REVIEWED BY
Yannis Skarpelos,
Panteion University, Greece

*CORRESPONDENCE
Divya Lakshmi S
✉ divyabalu19@gmail.com

RECEIVED 01 July 2025
ACCEPTED 29 September 2025
PUBLISHED 03 November 2025

CITATION
Lakshmi S D, Mohan V, Reeja PS, Alexander E, Varghese N and Kurian A (2025) Whose bias gets coded? Psychology's role in decolonizing AI. *Front. Psychol.* 16:1657227. doi: 10.3389/fpsyg.2025.1657227

KEYWORDS

facial features, gendered competence bias, DSM-based distress, Western norms, Western-centric cultural representation, proxies for risk, risk profiling, gang affiliation

Introduction

Who decides whose minds—and whose biases—get embedded in the algorithms that shape our world? The discipline of psychology, with its foundational assumptions and cultural biases, is deeply woven into artificial intelligence systems, influencing whose perspectives are valued and whose are marginalized. Bias in artificial intelligence is often described as a technical flaw, yet many of its origins are psychological—rooted in the constructs, methods, and epistemic values of psychology as a field. At the same time, bias also emerges from technical design choices, economic incentives, and governance structures. Our focus here is on psychology's distinctive role within this broader sociotechnical system (Ukanwa, 2024). This article argues that psychology both contributes to AI bias and holds unique potential to decolonize it through culturally inclusive design. To make this case, we first show how psychological constructs, often derived from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) contexts, have been embedded into AI in ways that amplify inequities. We then draw on decolonial traditions in psychology to propose standards and practices that could steer AI toward global representativeness, epistemic justice, and liberatory outcomes.

Algorithms inherit ways of thinking that reflect Western-centric models of behavior and identity (Rodier et al., 2023). Foundational critiques in AI ethics have already shown that models risk reproducing and amplifying social and cultural hierarchies, for instance uncover in the influential "stochastic parrots" argument that large-scale models inevitably encode representational harms (Bender et al., 2021).

As machine learning models increasingly rely on psychological constructs such as decision-making, affect detection, and personality assessment in simulating or forecasting human behavior, such constructs—frequently borrowed from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies—are universalized globally, often in spite of cultural specificity (Li et al., 2024; Peters and Carman, 2024). Recent empirical studies show that, in what ways language models possess not only linguistic but cultural and ideological biases as well, and we are forced to consider whose voices and perspectives they are representative of Santurkar et al. (2023). Such a procedure makes the AI technologies instruments for the communication not only of wisdom but ideological imperatives, and brings to the fore pertinent questions regarding whose psychology gets encoded and at whose expense (Shukla, 2025). Calls for more pluralistic models of AI alignment (Sorensen et al., 2024) are accurate where such a procedure exists, but a decolonial lens expands our eye to awareness that what gets encoded is not only behavior or cognitive processes but ideology, and brings with it the mandate for a psychological science realignment as a primary consideration in constructing ethically solidary, culturally inclusive technologies.

Cognitive bias: psychology's double-edged sword

Psychology did much in teaching us about bias, from the initial work of Kahneman and Tversky on heuristics to the most recent on implicit bias (Dominguez-Catena et al., 2025). Although this research has informed contemporary AI development, it has not proved sufficient to prevent the emergence and propagation of harmful biases in AI systems. We recognize, however, that bias is not always inherently negative; as recent work shows, certain forms of bias can be benign or even beneficial, while others entrench inequities (Fabi and Hagendorff, 2022; Waters and Honenberger, 2025). Our focus in this article is on those biases that reproduce systemic disadvantage and cultural exclusion.

A classic example is in facial recognition technology employed in Detroit, where in 2023 the system misclassified Black faces five to 10 times more often than white faces (Basheer, 2024). The imbalance is traced to training sets dominated by Caucasian facial features and to psychological assumptions—such as the idea that emotions are universally expressed and recognized across all cultures—that do not account for cultural and phenotypic differences (Domnich and Anbarjafari, 2021). The National Institute of Standards and Technology confirmed these findings, reporting racial disparities in 189 algorithms developed by 99 different companies (Grother et al., 2019).

Similarly, Amazon's defunct recruitment algorithm penalized resumes containing keywords such as "women's chess club" or references to all-female colleges. The system reproduced historical gender biases embedded in its training data, effectively privileging male-coded language and experiences. While not a psychological mechanism *per se*, this outcome parallels attribution errors studied in psychology, where competence is inferred through biased cues, and illustrates how data-driven systems can replicate structural inequities (Mahapatra and Mujtaba, 2019; Venkateshwaran, 2025).

These cases illustrate how psychological constructs, when uncritically embedded into AI, can entrench and even amplify social inequities. Table 1 summarizes additional real-world examples of psychological constructs in AI systems and their resulting sociocultural impacts.

Standards for bias in AI

The preceding examples underscore how psychological constructs, when uncritically implemented, amplify inequities. But this raises a critical question: is all bias inherently harmful? Some scholars argue that the goal of "bias-free AI" is both unattainable and misleading, since all models reflect values, assumptions, and standpoints (Fabi and Hagendorff, 2022). The key issue is not whether bias exists, but whether it is reflexively acknowledged, equitably distributed, and aligned with principles of justice.

From a decolonial perspective, the distinction lies between biases that "punch down" (reinforcing domination and exclusion) and those that "punch up" (resisting oppression and amplifying marginalized voices). Decolonial psychology traditions—from Fanon's analyses of colonial subjectivity (Fanon, 1967) to Martín-Baró's liberation psychology (Martín-Baró, 1994) and Indigenous frameworks of relationality emphasizing interconnectedness and collective wellbeing (Wilson, 2008)—remind us that partiality can serve emancipatory ends when it surfaces silenced knowledge systems rather than suppressing them.

This article therefore rejects the ideal of "bias-free AI" in favor of a normative standard: AI systems must disclose their epistemic commitments, embrace cultural pluralism, and actively counter oppressive hierarchies. In practice, this means building technologies that recognize the inevitability of bias, while steering it toward inclusivity, accountability, and justice.

Decolonizing psychology: the internal reckoning

The issue of bias in AI is not only technical but also connected to the long history of colonialism in psychology—pathologizing native cultures, medicalizing difference, and exporting Western diagnostic guidelines as if they were universal (Deva et al., 2025). Such epistemological biases are perpetually integrated into AI technologies. For example, the talk therapy conversational bots Wysa and Woebot, developed almost entirely from Western clinical databases, incorrectly interpreted culturally characteristic Indian expressions. A 2024 study identified that Indian users' statements such as "family pressure is my karma" were pathologized as depressive symptoms, deserving of unsuitable therapeutic interventions (Viberg et al., 2023). Generative models such as Midjourney and DALL-E also incorrectly represented Indian subcultures: a 2024 University of Pennsylvania study noted that the prompt "Indian family dinner" produced poverty-related images 73% of the time, while "American family dinner" produced middle-class settings.

These are described as psychological assumptions because they draw directly from psychological constructs—such as the universality of diagnostic categories (e.g., DSM-based criteria for distress), standardized measures of emotion, and Western notions of family and identity—that have historically been treated as objective and culture-free. When transferred into AI systems, such constructs enact a second-order colonization: exporting Western norms while erasing non-Western modalities of being under the guise of objectivity and universality (Ghosh et al., 2024; Kumar et al., 2025).

Addressing injustice through decolonial AI practices

Decolonial solutions in practice

Psychology can help repair these damages by fostering epistemic pluralism, co-design, and critical reflexivity. One strong

TABLE 1 Examples of psychological bias in AI and their impacts.

Case study/Technology	Psychological construct embedded	Cultural/ Demographic impacted	Key consequence/Harm	Citation(s)
Detroit facial recognition	Universality of emotion/facial features	Black Americans	The system produced misidentification rates five to 10 times higher for Black individuals, resulting in wrongful arrests.	Basheer, 2024 Domnich and Anbarjafari, 2021; National Institute of Standards and Technology, 2022
Amazon recruitment algorithm	Gendered competence bias	Women applicants	Penalized resumes mentioning "women's," underrepresentation	Mahapatra and Mujtaba, 2019; Venkateshwaran, 2025
Indian mental health chatbots (Wysa, Woebot)	DSM-based distress, Western norms	Indian users	Misinterpretation of local idioms, inappropriate recommendations	Viberg et al., 2023; Deva et al., 2025
Midjourney/DALL-E Generative AI	Western-centric cultural representation	Indian subcultures	73% poverty imagery for "Indian family dinner" prompts	Ghosh et al., 2024; Kumar et al., 2025
Dutch tax authority fraud algorithm	Proxies for risk (dual nationality, income)	Ethnic minorities	20,000 families wrongly accused, 1,000 children in foster care	Alba, 2024
Chicago predictive policing	Risk profiling, "gang affiliation"	Black neighborhoods	32% higher arrests despite equal crime rates	Khan and Ewuoso, 2024

example comes from South Africa, where researchers collaborated with Zulu communities to co-design a tuberculosis (TB) diagnosis algorithm. In Zulu, the idiom "isifuba sibuhlungu" literally means "the chest is painful," but it carries cultural and linguistic nuances that early English-trained diagnostic systems failed to interpret correctly. Because these AI models were trained only on biomedical descriptions in English, they initially misclassified the condition, lowering diagnostic accuracy to 68%. By incorporating Zulu terminology and involving community members in the design process, accuracy improved dramatically to 92%. This shows that co-design not only raises technical precision but also affirms the legitimacy of local knowledge in medical AI.

There is a similar lesson from Aotearoa/New Zealand, in which Māori researchers established the Te Hiku voice model as an embodiment of the speaking voice of te reo Māori. Off-the-shelf commercial speech-recognition software had an error rate of as high as 47% among Māori speakers directly because they were excluded from international training datasets. By collating and controlling their own data sets, Māori communities brought errors to near zero while keeping their voices and the way they are stored in sovereign hands. The project proves that Indigenous-driven design is both feasible in terms of system performance increase and in terms of cultural sovereignty protection.

These exemplars conform with the new GPA guidelines (Global Psychology Alliance, 2023), that emphasize three practices: cultural disclosure, cross-validation, and site-based audits. Cultural disclosure obliges researchers to be transparent in marking the cultural heritage of their data so that Western constructs will cease being presented as universal. Cross-validation over three or more settings demands that psychological constructs and AI models be proved in diverse cultural settings rather than in WEIRD populations so that they can be applied more widely. Site-based audits demand independent, often community-based examination of psychological research and applications of AI so that ethical propriety as well as cultural sensibility may be established.

Applied to AI, these standards mean that therapeutic chatbots, diagnostic algorithms, or predictive models should (1) disclose the cultural provenance of their training data, (2) undergo validation in multiple culturally distinct populations, and (3) be audited by independent panels that include community stakeholders. Together, these measures shift psychology from passive collaborator in bias to active ethical design architect, embedding justice and cultural pluralism into the development of AI (Masaka, 2019; Ofosu-Asare, 2024).

Rewriting the code of psychology

The question of whose bias gets coded necessarily opens onto the complicity of psychology in the empowerment of systems that disguise ideology as objectivity. The examples—from the Dutch tax algorithm's racial bias through Chicago's predictive policing—are used here just to illustrate how psychological inventions, disguised under the pretense of universal laws, have themselves been used to encode the colonial relations of power in AI. Not the predestined product of technology but the product of political and disciplinary choices, they may be undone through the decentralized production of knowledge through both cross-cultural validation as well as community co-design. Cross-cultural validation and community co-design are related but distinct processes: validation demonstrates that a model's performance is not WEIRD-restricted, while co-design ensures the model is legitimate, safe, and valuable in a specific cultural setting. Tensions will inevitably arise—for example, when a community prefers a design that is not yet cross-culturally validated, or when a validated system conflicts with local norms. In such cases, we recommend a dual-criterion protocol weighing generalizability against local acceptability: deployments are only to proceed when both a generalizability threshold (G) and a locally set acceptability threshold (L) are met. Where threshold values are different, deployment is either to be limited to local scale with explicit labeling (L≥L*, G<G*) or deferred until redesign and governance protection are in place (G≥G*, L<L*). It thus makes explicit why co-design is necessary, in what respects it is different from validation, and in what respects such conflicts are responsibly to resolve.

Epistemic responsibility must draw sustenance also from independent data and algorithmic audits. Such audits must take place under the aegis of cross-disciplinary working groups of academic specialists, civic societies, and citizens from communities most at risk of the effects of AI. Their remit is in technical proficiency as well as in representational democracy, such that fairness, cultural compatibility, and transparency must be taken into account in many different lexicons. Justice must become an organizing principle both in the generation and in use of AI. Psychology is at a fork in the road at present: it can become an architecture of exclusion, or become the scaffold for liberation through the generation of culturally compatible and ethically informed systems of AI. The code, and the effects of code, is ours to rewrite.

Author contributions

DL: Resources, Formal analysis, Validation, Writing – review & editing, Conceptualization, Writing – original draft. VM: Conceptualization, Writing – original draft, Writing – review & editing, Formal analysis. PR: Writing – original draft, Writing – review & editing, Conceptualization. EA: Formal analysis, Writing – original draft, Writing – review & editing. NV: Writing – review & editing, Conceptualization, Writing – original draft. AK: Writing – review & editing, Formal analysis, Writing – original draft.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alba, J. T. (2024). Insights into algorithmic decision-making systems via a decolonial-intersectional lens: a cross-analysis case study. *Digit. Soc.* 3:58. doi: 10.1007/s44206-024-00144-9

Basheer, I. (2024). Bias in the algorithm: issues raised due to use of facial recognition in India. *J. Dev. Policy Pract.* 10, 61–79. doi: 10.1177/24551333241283992

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). "On the dangers of stochastic parrots: can language models be too big?," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event), 610–623.

Deva, R., Ramani, D., Divate, T., Jalota, S., and Ismail, A. (2025). "Kya family planning after marriage hoti hai?: integrating cultural sensitivity in an LLM chatbot for reproductive health," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems,* 1–23.

Dominguez-Catena, I., Paternain, D., Galar, M., Defrance, M., Buyl, M., and Bie, T. (2025). Biased heritage: how datasets shape models in facial expression recognition. *arXiv. arXiv:2503.03446.* doi: 10.48550/arXiv.2503.03446

Domnich, A., and Anbarjafari, G. (2021). Responsible AI: Gender bias assessment in emotion recognition. *arXiv. arXiv:2103.11436.* doi: 10.48550/arXiv.2103.11436

Fabi, S., and Hagendorff, T. (2022). Why we need biased AI: How including cognitive and ethical machine biases can enhance AI systems. *arXiv* [Preprint]. *arXiv:2203.09911.* Available online at: https://arxiv.org/abs/2203.09911

Fanon, F. (1967). *Black Skin, White Masks.* New York, NY: Grove Press.

Ghosh, S., Venkit, P. N., Gautam, S., Wilson, S., and Caliskan, A. (2024). *Do generative AI models output harm while representing non-Western* cultures: evidence from a community-centered approach. *arXiv* [Preprint]. *arXiv:2407.14779.* doi: 10.48550/arXiv.2407.14779

Global Psychology Alliance (2023). *Guidelines for cultural responsiveness and equity in psychological science.* International Union of Psychological Science. Available online at: https://globalpsychologyalliance.org/resources/guidelines-cultural-responsiveness-equity

Grother, P., Ngan, M., and Hanaoka, K. (2019). *Face recognition vendor test part 3: Demographic effects (NIST IR 8280; p. NIST IR 8280).* Gaithersburg, MD: National Institute of Standards and Technology.

Khan, M., and Ewuoso, C. (2024). Epistemic (in)justice, social identity and the black box problem in patient care. *Med. Health Care Philos.* 27, 227–240. doi: 10.1007/s11019-024-10194-y

Kumar, A., Chakraborty, A., and Chhikara, G. (2025). *Through the prism of* culture: Evaluating LLMs' understanding of Indian subcultures and traditions. *arXiv* [Preprint]. *arXiv:2303.17548.* doi: 10.48550/arXiv.2501.16748

Li, S., Yuan, H., Hu, X., Che, Z., Zhang, Y., and Luo, S. (2024). The high dimensional psychological profile and cultural bias of ChatGPT. *Sci. China Technol. Sci.* 68:1820401. doi: 10.1007/s11431-025-2934-8

Mahapatra, N., and Mujtaba, D. (2019). "Ethical considerations in AI-based recruitment," in *2019 IEEE International Symposium on Technology and Society (ISTAS)* (Medford, MA: IEEE), 1–7.

Martín-Baró, I. (1994). *Writings for a Liberation Psychology.* Cambridge, MA: Harvard University Press.

Masaka, D. (2019). Attaining epistemic justice through transformation and decolonisation of education curriculum in Africa. *Afr. Ident.* 17, 298–309. doi: 10.1080/14725843.2019.1681259

National Institute of Standards and Technology (2022). *NIST conference papers fiscal year 2019 (NIST Special Publication 1276v1)*. doi: 10.6028/nist.sp.1276v1

Ofosu-Asare, Y. (2024). Cognitive imperialism in artificial intelligence: counteracting bias with indigenous epistemologies. *AI Soc.* 40, 3045–3061. doi: 10.1007/s00146-024-02065-0

Peters, U., and Carman, M. (2024). *Cultural bias in explainable AI* research: a systematic analysis. *J. Artif. Intell. Res.* 79, 971–1000. doi: 10.1613/jair.1.14888

Rodier, M., Bernault, C., André, J., Delmas, A., Juan, S., and Kaadoud, I. C. (2023). "Assessing the impact of cognitive biases in AI project development," in *Artificial Intelligence in HCI. HCII 2023. Lecture Notes in Computer Science*, eds. H. Degen and S. Ntoa (Cham: Springer), 401–420.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. (2023). Whose opinions do language models reflect? *arXiv. arXiv:2303.17548.* doi: 10.48550/arXiv.2303.17548

Shukla, N. (2025). Investigating AI systems: examining data and algorithmic bias through hermeneutic reverse engineering. *Front. Commun.* 10:1380252. doi: 10.3389/fcomm.2025.1380252

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., et al. (2024). A roadmap to pluralistic alignment. *arXiv. arXiv:2402.05070.* doi: 10.48550/arXiv.2402.05070

Ukanwa, K. (2024). Algorithmic bias: social science research integration through the 3-D dependable AI framework. *Curr. Opin. Psychol.* 58:101836. doi: 10.1016/j.copsyc.2024.101836

Venkateshwaran, G. (2025). *Artificial intelligence* in HR: transforming recruitment and selection in IT industry. *J. Inf. Syst. Eng. Manag.* 10, 38–45. doi: 10.52783/jisem.v10i17s.2705

Viberg, O., Kizilcec, R., Baker, R., and Tao, Y. (2023). Cultural bias and cultural alignment of large language models. *PNAS Nexus* 3:pgae346. doi: 10.1093/pnasnexus/pgae346

Waters, G., and Honenberger, P. (2025). AI biases as asymmetries: a review to guide practice. *Front. Big Data* 8:1532397. doi: 10.3389/fdata.2025.1532397

Wilson, S. (2008). *Research is Ceremony: Indigenous Research Methods*. Oxford: Fernwood Publishing.