Provocation: Who benefits from "inclusion" in Generative AI?

Samantha Dalal*
University of Colorado Boulder
samantha.dalal@colorado.edu

Siobhan Mackenzie Hall*
University of Oxford
siobhan.hall@nds.ox.ac.uk

Nari Johnson*
Carnegie Mellon University
narij@andrew.cmu.edu

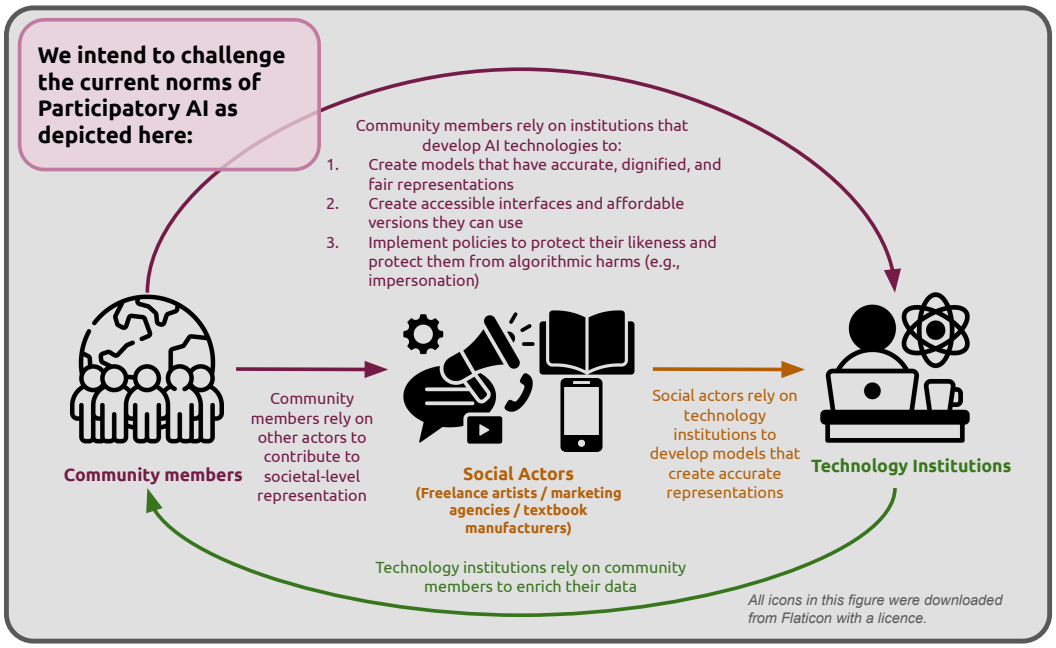

Figure 1: We map the dependencies that often exist between key stakeholder groups: (1) members of marginalized communities, (2) other social actors who use GenAI tools, and (3) institutions (such as private technology companies or academic universities) that develop GenAI models, in participatory engagements to improve how marginalized communities are represented by GenAI. In this paper, we discuss how these dependencies pose barriers for marginalized communities to realize the benefits of improved AI models.

Socially marginalized groups disproportionately experience *representational harms* caused by generative AI systems [1, 2, 3, 4, 5, 6, 7]. For example, popular text-to-image (T2I) models have been shown to generate inaccurate, culturally misrepresentative, and insensitive depictions of racial and ethnic minorities [1], people with disabilities [5], and foods from the African continent [4]. As the AI community begins to acknowledge the limits of Internet-scraped datasets and the narrowing of

*Equal contribution.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

values and perspectives involved in AI design procedures [8, 9, 10, 11, 12], there is an undeniable need to move towards including more community expertise and input. However, participation from community members runs the risk of being extractive, as many participatory engagements involve exposing participants to psychologically harmful or demeaning AI-generated content [2, 4, 5, 13]. More broadly, expertise is shifted from marginalized communities to AI model owners without commensurate structures for continued community agency and ownership over outputs [14, 15, 16].

In this provocation, we argue that dominant structures of community participation in AI development and evaluation are not explicit enough about the benefits and harms that members of socially marginalized groups may experience as a result of their participation [17]. Participation is increasingly motivated by trickle-down effect logics: model improvement will address stereotypes and help preserve material culture [2, 3, 4]. However, the potential for extractive and exploitative practices in participation is not necessarily given the same consideration. We are concerned that the claim that community members will be better off as a result of their participation is empty, given the immensity of systemic change that is needed *as well*. We present a speculative case study [18, 19, 20], based on our collective experiences doing community-engaged research in AI, to interrogate the promises AI developers make to members of marginalized groups, and itemize the barriers that realistically need to be overcome for the proposed benefits to marginalized communities to be realized.

Speculative Case Study: How do community members realize benefits & harms from improved GenAI performance? Researchers do not yet fully understand how to leverage technical machine learning capabilities to improve the representation of marginalized communities in GenAI models [21, 22, 23, 24]. Thus, we do not always know whether incorporating community feedback during the development process will necessarily lead to an "improved" AI model. Regardless, we believe it is important to investigate the key premise motivating participatory approaches to AI: *(How) do improved representations in GenAI models benefit members of marginalized groups*?

To interrogate whom dominant structures of community participation in GenAI development benefit, we present a hypothetical scenario where a technology company invites Vietnamese community members to participate in the development and evaluation of a T2I system. The case study that we present is meant to be an abstraction of higher-level themes that we observed over our experiences working as AI researchers within industry and academic contexts. We use this grounding context to trace the flow of potential benefits and harms between different groups of stakeholders.

Scenario: Thuy, a cultural preservation activist in Vietnam, is invited to participate in a technology company's AI data enrichment initiative where community members label photographs of Vietnamese cultural artifacts for AI training and evaluation. The company aims to improve the depiction of "Majority World" cultures [25] in T2I systems to support improved *quality-of-service* [26], and Thuy is excited to partake in this effort to improve Vietnamese representation in global media. The company believes that improved *quality-of-service* can help Vietnamese users create images that accurately reflect their culture, for personal projects or educational purposes. Other social actors, such as marketing agencies, textbook publishers, and freelance artists, can also benefit from generating inclusive and accurate media. The technology company adopts a common structure of participation [27] to engage with Thuy and other Vietnamese community advocates: they provide participants with one-time compensation for providing data and expertise. The company has not explored paths for participant ownership or control over data or AI models that are created as a result of the engagement.

(How) can community members benefit when they are end-users? Thuy and other members of her community may face *financial barriers* in realizing the benefits of improved quality of service in T2I models. While Thuy is provided with one-time compensation for her participation, she does not continue to financially benefit from the future use of her data or the AI models it was used to improve. The company can improve its T2I offerings and monetize its competitive advantage by putting its services behind a paywall. Thuy's peers and other members can now use the model to generate accurate depictions of their likeness, but must navigate the company's paywall structures. Due to a lack of ownership over their data and resulting AI models, community ambassadors like Thuy are *sold back models with improvements that would not have been possible without their labor*.

Beyond navigating paywalls, Thuy and other community members face additional barriers in *accessing* and using the company's models. For example, for many marginalized communities, model access can be complicated by several potential issues such as a lack of reliable Internet connectivity [4] and inaccessible user interfaces [28, 29]. Thus, *Thuy is unlikely to be able to realize the benefits of the model's improved performance as an end-user if she cannot reliably access and navigate its interface*.

(How) can community members benefit when they are not end-users? Thuy may face *socio-political barriers* in realizing indirect benefits resulting from social actors using T2I models as end-users to create images of the Vietnamese community. Many researchers have argued that due to the increasing prevalence of AI-generated media, GenAI systems will shape *societal representation* [5, 6, 30, 31, 32] and thus precipitate change in societal attitudes towards marginalized communities. For example, a marketing firm may use the GenAI model to create images for an ad campaign that depicts Vietnamese people and culture, which is then seen by millions of people.

However, media studies scholars have identified that representation in media *alone* will not result in direct change to material circumstances for marginalized communities [33, 34, 35]. The political economy of the media ecosystem, including industry logics and financial incentives, dictates the kinds of media that are produced [36]. Thus, Thuy is unlikely to realize the benefits of social actors using T2I models to create images of her community *unless* social, political, and economic conditions all align to transform visibility into political power.

Harms marginalized groups can experience as a result of their participation Increased visibility in AI-generated media may make marginalized communities susceptible to a wide and emerging range of AI-mediated harms [26, 37]. For example, as social actors (*e.g.,* textbook companies) realize that they can use AI to generate accurate depictions of Vietnamese culture, they may no longer consult or compensate Vietnamese community members, resulting in further financial and social marginalization [38, 39, 40]. Other actors can exploit improved representations of marginalized communities to inflict harm such as impersonation, misinformation, or the creation of violent/NSFW content [41, 42]. Thus, *members of marginalized communities rely on technology institutions to implement effective policies to protect their likeness*. While the technology company could implement mitigation steps (*e.g.,* access restrictions or usage licenses [43, 44]) to prevent misuse, they may find them at odds with their profit motives.

Implications While the details of this scenario were speculative, the discussed model of participatory engagement as one-time consultation illustrates the reality of how technology institutions and academic researchers often engage socially marginalized communities in AI development today [17, 27, 45]. Thus, we urge the broader AI community, including those who construct or participate in participatory engagements, to critically evaluate whether these dominant structures of participation do in fact yield their intended benefits *for marginalized communities*. AI researchers and industry actors who are conducting participatory engagements with marginalized communities should be more transparent to participants and the community about the accessibility of benefits to participants and the contingencies upon which these benefits rely. In Appendix A, we pose future directions and highlight promising examples towards restructuring participation beyond consultation, and towards supporting meaningful community ownership, participation, and power over AI.

1 Broader impact statement

As discussed in our provocation, we believe that participatory engagements with socially marginalized groups are critical to the broader field of AI and machine learning. We urge researchers to ask themselves how communities whose participation we solicit can benefit from improved model performance. This critical self-reflection requires that researchers map out both the direct benefits participants could experience as end-users and the indirect benefits participants could realize as a result of other social actors using AI systems developed with community input. Understanding how such participatory engagements can be structured is of timely importance given the rapidly advancing capabilities of generative AI; new regulatory and policy requirements that require consultation with impacted groups [46, 47]; and to combat the increasing centralization of power in who has a say in AI's increasing influence on society [16].

2 Limitations

We acknowledge the limitations of our analysis, which is centered around a speculative case study with imagined actors. The barriers to realizing benefits from AI that we surfaced in our case study were based on this speculative context informed by our past experiences (*e.g.,* the communities we are members of, or have engaged in research with before) and our positionality as AI researchers. Future work can engage more deeply in analyzing real-world examples of participatory engagements with socially marginalized groups, and understanding how barriers participants face when realizing benefits vary across shared identities and contexts.

In this short piece, we briefly sketch the "dominant structures" of participation [27] in GenAI evaluation, our concerns with these structures, and potential paths forward. In doing so, our goal is *not* to critique the premise that socially marginalized communities should be involved in AI development and evaluation; or that existing participatory efforts should not be pursued. Rather, we remain hopeful that more deeply interrogating how participation is structured can lead to more empowering and constructive ways of engaging socially marginalized communities. Deeper engagement beyond what we could do within this workshop paper contribution is required to understand how structures of participation in GenAI development and evaluation can be shaped to support the equitable distribution of benefits and power among stakeholders.

Acknowledgments and Disclosure of Funding

We thank Michael Madaio, Calvin Liang, Michael Feffer, and the anonymous reviewers at the NeurIPS 2024 EvalEval Workshop for offering feedback on this work. NJ acknowledges support from the NSF (IIS2040929 and IIS2229881) and the Block Center for Technology and Society at CMU. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of the National Science Foundation and other funding agencies.

References

[1] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1493–1504, New York, NY, USA, 2023. Association for Computing Machinery.

[2] Melissa Hall, Samuel J. Bell, Candace Ross, Adina Williams, Michal Drozdzal, and Adriana Romero Soriano. Towards geographic inclusion in the evaluation of text-to-image models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 585–601, New York, NY, USA, 2024. Association for Computing Machinery.

[3] Agrima Seth, Sanchit Ahuja, Kalika Bali, and Sunayana Sitaram. Dosa: A dataset of social artifacts from different indian geographical subcultures, 2024.

[4] Jabez Magomere, Shu Ishida, Tejumade Afonja, Aya Salama, Daniel Kochin, Foutse Yuehgoh, Imane Hamzaoui, Raesetje Sefala, Aisha Alaagib, Elizaveta Semenova, Lauren Crais, and Siobhan Mackenzie Hall. You are what you eat? feeding foundation models a regionally diverse food dataset of world wide dishes. 2024.

[5] Kelly Avery Mack, Rida Qadri, Remi Denton, Shaun K. Kane, and Cynthia L. Bennett. "they only care to show us the wheelchair": disability representation in text-to-image ai models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[6] Jared Katzman, Angelina Wang, Morgan Scheuerman, Su Lin Blodgett, Kristen Laird, Hanna Wallach, and Solon Barocas. Taxonomizing and measuring representational harms: A look at image tagging, 2023.

[7] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. The problem with bias: From allocative to representational harms in machine learning. In *SIGCIS conference paper*, 2017.

[8] Abeba Birhane, Sepehr Dehdashtian, Vinay Prabhu, and Vishnu Boddeti. The dark side of dataset scaling: Evaluating racial classification in multimodal models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1229–1244, New York, NY, USA, 2024. Association for Computing Machinery.

[9] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. The values encoded in machine learning research, 2022.

[10] Rachel Hong, William Agnew, Tadayoshi Kohno, and Jamie Morgenstern. Who's in and who's out? a case study of multimodal clip-filtering in datacomp, 2024.

[11] Terrance DeVries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone?, 2019.

[12] Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In Sunipa Dev, Vinodkumar Prabhakaran, David Adelani, Dirk Hovy, and Luciana Benotti, editors, *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[13] Alice Qian Zhang, Ryland Shaw, Jacy Reese Anthis, Ashlee Milton, Emily Tseng, Jina Suh, Lama Ahmad, Ram Shankar Siva Kumar, Julian Posada, Benjamin Shestakofsky, Sarah T. Roberts, and Mary L. Gray. The human factor in ai red teaming: Perspectives from social and collaborative computing. *arXiv preprint arXiv:2407.07786*, 2024.

[14] Jennifer Pierre, Roderic Crooks, Morgan Currie, Britt Paris, and Irene Pasquetto. Getting ourselves together: Data-centered participatory design research & epistemic burden. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–11, Yokohama Japan, May 2021. ACM.

[15] Victor Ojewale, Ryan Steed, Briana Vecchione, Abeba Birhane, and Inioluwa Deborah Raji. Towards ai accountability infrastructure: Gaps and opportunities in ai audit tooling, March 2024.

[16] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. Power to the people? opportunities and challenges for participatory ai. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, New York, NY, USA, 2022. Association for Computing Machinery.

[17] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. The participatory turn in ai design: Theoretical foundations and the current state of practice, 2023.

[18] Casey Fiesler. Innovating like an optimist, preparing like a pessimist: Ethical speculation and the legal imagination. *Colo. Tech. LJ*, 19:1, 2021.

[19] Nina Bozic Yams and Álvaro Aranda Muñoz. Poetics of future work: Blending speculative design with artistic methodology. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA, 2021. Association for Computing Machinery.

[20] Shamika Klassen and Casey Fiesler. The stoop: speculation on positive futures of black digital spaces. *Proceedings of the ACM on Human-Computer Interaction*, 7(GROUP):1–24, 2023.

[21] Daniela Massiceti, Camilla Longden, Agnieszka Słowik, Samuel Wills, Martin Grayson, and Cecily Morrison. Explaining clip's performance disparities on data from blind/low vision users, 2024.

[22] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models, 2023.

[23] Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 786–808, New York, NY, USA, 2023. Association for Computing Machinery.

[24] Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation, 2023.

[25] Shahidul Alam. Majority world: Challenging the west's rhetoric of democracy. *Amerasia Journal*, 34:87–98, 01 2008.

[26] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction, 2023.

[27] Harini Suresh, Emily Tseng, Meg Young, Mary Gray, Emma Pierson, and Karen Levy. Participation in the age of foundation models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1609–1621, New York, NY, USA, 2024. Association for Computing Machinery.

[28] Aleksi Väisänen. Guidelines supported evaluation of user interfaces with generative ai. 2024.

[29] Maitraye Das, Alexander J. Fiannaca, Meredith Ringel Morris, Shaun K. Kane, and Cynthia L. Bennett. From provenance to aberrations: Image creator and screen reader user perspectives on alt text for ai-generated images. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[30] Tarleton Gillespie. Generative ai and the politics of visibility. *Big Data & Society*, 11(2):20539517241252131, June 2024.

[31] Safiya Umoja Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, New York, 2018.

[32] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[33] Herman Gray. Subject (ed) to recognition. *American Quarterly*, 65(4):771–798, 2013.

[34] Kristen J. Warner. In the time of plastic representation. *Film Quarterly*, 71(2):32–37, 2017.

[35] Anamik Saha. Beards, scarves, halal meat, terrorists, forced marriage': television industries and the production of 'race. *Media, Culture & Society*, 34(4):424–438, 2012.

[36] Adrienne Shaw and Katherine Sender. Queer technologies: Affordances, affect, ambivalence, 2016.

[37] Atli Sigurgeirsson and Eddie L. Ungless. Just because we camp, doesn't mean we should: The ethics of modelling queer voices, 2024.

[38] Riddhi Setty. Ai threatens to push human fashion models out of the picture, January 2024. Accessed: 2024-09-10.

[39] William Agnew, A. Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R. McKee. The illusion of artificial inclusion. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, volume 35 of *CHI '24*, page 1–12. ACM, May 2024.

[40] Cedric Deslandes Whitney and Justin Norman. Real risks of fake data: Synthetic data, diversity-washing and consent circumvention. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1733–1744, New York, NY, USA, 2024. Association for Computing Machinery.

[41] Felipe Romero Moreno. Generative ai and deepfakes: a human rights approach to tackling harmful content. *International Review of Law, Computers & Technology*, pages 1–30, 2024.

[42] Aurélie Petit. The limits of zero tolerance policies for animated pornographic media. *Porn Studies*, 2024. Special Issue 'Artificial Intelligence, Pornography, and Sex Work' (forthcoming).

[43] Paul T. Brown, Daniel Wilson, Kiri West, Kirita-Rose Escott, Kiya Basabas, Ben Ritchie, Danielle Lucas, Ivy Taia, Natalie Kusabs, and Te Taka Keegan. Māori algorithmic sovereignty: idea, principles, and use, 2023.

[44] The Data Science Law Lab. Licensing african datasets, September 2024. Accessed: 2024-09-10.

[45] Meg Young, Upol Ehsan, Ranjit Singh, Emnet Tafesse, Michele Gilman, Christina Harrington, and Jacob Metcalf. Participation versus scale: Tensions in the practical demands on participatory ai. *First Monday*, 29(4), Apr. 2024.

[46] Shiming Hu and Yifan Li. Policy interventions and regulations on generative artificial intelligence: Key gaps and core challenges. In *Proceedings of the 25th Annual International Conference on Digital Government Research*, dg.o '24, page 1034–1036, New York, NY, USA, 2024. Association for Computing Machinery.

[47] Philipp Hacker, Andreas Engel, and Marco Mauer. Regulating chatgpt and other large generative ai models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1112–1123, New York, NY, USA, 2023. Association for Computing Machinery.

[48] Kimberly A Christen. Does information really want to be free? indigenous knowledge systems and the question of openness. *International Journal of Communication*, 6:24, 2012.

[49] Nicholas Vincent, Hanlin Li, Nicole Tilly, Stevie Chancellor, and Brent Hecht. Data leverage: A framework for empowering the public in its relationship with technology companies. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 215–227, 2021.

[50] Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, et al. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619*, 2024.

[51] Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Hassan Muhammad, Salomon Kabongo, Salomey Osei, Sackey Freshia, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa, Mofe Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Jane Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkabir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Espoir Murhabazi, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Emezue, Bonaventure Dossou, Blessing Sibanda, Blessing Itoro Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. Participatory research for low-resourced machine translation: A case study in african languages, 2020.

[52] David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, Frederico Belcavello, Ganzorig Batnasan, Gisela Vallejo, Grainne Caulfield, Guido Ivetta, Haiyue Song, Henok Biadglign Ademtew, Hernán Maina, Holy Lovenia, Israel Abebe Azime, Jan Christian Blaise Cruz, Jay Gala, Jiahui Geng, Jesus-German Ortiz-Barajas, Jinheon Baek, Jocelyn Dunstan, Laura Alonso Alemany, Kumaranage Ravindu Yasas Nagasinghe, Luciana Benotti, Luis Fernando D'Haro, Marcelo Viridiano, Marcos Estecha-Garitagoitia, Maria Camila Buitrago Cabrera, Mario Rodríguez-Cantelar, Mélanie Jouitteau, Mihail Mihaylov, Mohamed Fazli Mohamed Imam, Muhammad Farid Adilazuarda, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Naome Etori, Olivier Niyomugisha, Paula Mónica Silva, Pranjal Chitale, Raj Dabre, Rendi Chevi, Ruochen Zhang, Ryandito Diandaru, Samuel Cahyawijaya, Santiago Góngora, Soyeong Jeong, Sukannya Purkayastha, Tatsuki Kuribayashi, Teresa Clifford, Thanmay Jayakumar, Tiago Timponi Torrent, Toqeer Ehsan, Vladimir Araujo, Yova Kementchedjhieva, Zara Burzo, Zheng Wei Lim, Zheng Xin Yong, Oana Ignat, Joan Nwatu, Rada Mihalcea, Thamar Solorio, and Alham Fikri Aji. Cvqa: Culturally-diverse multilingual visual question answering benchmark, 2024.

[53] Hellina Hailu Nigatu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury. The zeno's paradox of 'low-resource' languages. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17753–17774, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[54] Atnafu Lambebo Tonja, Bonaventure F. P. Dossou, Jessica Ojo, Jenalea Rajab, Fadel Thior, Eric Peter Wairagala, Anuoluwapo Aremu, Pelonomi Moiloa, Jade Abbott, Vukosi Marivate, and Benjamin Rosman. Inkubalm: A small language model for low-resource african languages, 2024.

[55] Karen Hao. A new vision of artificial intelligence for the people, 2022.

[56] Asmelash Teka Hadgu and Paul Azunre and Timnit Gebru. Combating harmful hype in natural language processing, 2023. Accessed: 2024-09-10.

[57] Danish Contractor, Daniel McDuff, Julia Katherine Haines, Jenny Lee, Christopher Hines, Brent Hecht, Nicholas Vincent, and Hanlin Li. Behavioral use licensing for responsible ai. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 778–788. ACM, June 2022.

[58] Daniel McDuff, Tim Korjakow, Scott Cambo, Jesse Josua Benjamin, Jenny Lee, Yacine Jernite, Carlos Muñoz Ferrandis, Aaron Gokaslan, Alek Tarkowski, Joseph Lindley, A. Feder Cooper, and Danish Contractor. On the standardization of behavioral use clauses and their adoption for responsible licensing of ai, 2024.

[59] C Okorie and M Omino. Licensing african datasets, 2024.

[60] Junwei Deng, Shiyuan Zhang, and Jiaqi Ma. Computational copyright: Towards a royalty model for music generative ai, 2024.

[61] Eric P.S. Baumer and M. Six Silberman. When the implication is not to design (technology). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, page 2271–2274, New York, NY, USA, 2011. Association for Computing Machinery.

[62] Noura Howell, Audrey Desjardins, and Sarah Fox. Cracks in the success narrative: Rethinking failure in design research through a retrospective trioethnography. *ACM Trans. Comput.-Hum. Interact.*, 28(6), November 2021.

[63] Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. Assembling accountability: algorithmic impact assessment for the public interest, 2021.

[64] Tzu-Sheng Kuo, Hong Shen, Jisoo Geum, Nev Jones, Jason I. Hong, Haiyi Zhu, and Kenneth Holstein. Understanding frontline workers' and unhoused individuals' perspectives on ai used in homeless services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[65] Alicia DeVrio, Motahhare Eslami, and Kenneth Holstein. Building, shifting, & employing power: A taxonomy of responses from below to algorithmic harm. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1093–1106, 2024.

[66] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, 22(2014):4349–4357, 2014.

[67] Alice Qian Zhang, Judith Amores, Mary L. Gray, Mary Czerwinski, and Jina Suh. Aura: Amplifying understanding, resilience, and awareness for responsible ai content work, 2024.

[68] Karen Hao and Deepa Seetharaman. Cleaning up chatgpt takes heavy toll on human workers, 2023.

A Imagining paths forward

What sources of inspiration can researchers or facilitators of participatory AI initiatives turn to in their pursuit of more equitable engagement practices?

In this section, we share several resources and directions for paths forward. We do not aim to be comprehensive; rather, we highlight a few initiatives that propose alternatives to dominant participation structures. We loosely organize our discussion under two motivating questions. First, we look at theories from adjacent fields that, while not specifically about AI, provide valuable insights into participation and power. Second, we examine current efforts aimed at disrupting dominant structures and creating alternative modes of engagement that benefit marginalized communities.

Are there theories from literature or other forms of community-based knowledge that can inform paths forward for AI? Participation can be extractive. Communities can lose control over how their data is used and shared once it is collected for model training, fail to be credited for their contributions to model development, or not be properly compensated for their knowledge. Below we list some resources from Indigenous data studies, dataset development, and critical data studies that identify how researchers can respect communities' preferences around data sharing.

- Christen [48] identifies how some Indigenous communities have cultural norms for sharing certain types of data based on social relationships to the data artifact. AI researchers developing datasets of cultural artifacts with Indigenous communities can do work to first understand what cultural artifacts they are collecting that may have specific protocols for sharing. Researchers can then inform communities about the limitations of restricting access to images when developing and deploying GenAI models so communities can exert more informed consent.
- Vincent et al. [49] conceptualize the power that contributors hold over models as data leverage. Contributors to datasets can exert power over model development and performance by reducing, stopping, redirecting, or manipulating their data. Data leverage makes explicit technology companies' *dependence* on marginalized communities to improve their models' performance. Communities therefore have a significant amount of *leverage* to share the terms of their future inclusion in AI development. AI researchers should consider explaining to contributors the leverage they hold over the model development process as they address fair compensation for dataset contributions. Doing so could provide contributors with a way to conceptualize the value of their data and allow them to more critically assess the remuneration they are being offered for participating and the terms of their participation. In addition, AI researchers could use data leverage to calculate more accurate estimates of financial remuneration to contributors: How much would they be willing to pay to avoid contributors using their leverage to disrupt their model?

Are there example community collaborations that offer alternative models on how to structure participation in AI development/evaluation? Past scholarship [17, 27] has demonstrated how many "participatory AI" engagements are limited to consultation and inclusion (*e.g.,* collecting data from participants to enrich models), without granting participants meaningful opportunities for *ownership and control* over the resulting datasets and models. While participants may be able to give input on how they think the model should behave, ultimately, "participants have little say regarding the model's impact in the world: whether it is developed, what other data it is trained on, what it may be used for, or if and how it should be deployed" [27]. We identify some resources where researchers and communities have been developing alternative models to structure more equitable community engagement in AI development that ensures that participants have a meaningful say over model development **and** deployment.

1. *Alternative models of acknowledgment for participation.* Singh et al. [50] develop protocols for recognizing community contribution to AI development by operationalizing a broad definition of authorship for academic papers. Similar initiatives have also been led or adopted by other community-driven AI initiatives [51, 4, 52]. Papers are a valuable currency for visibility and recognition in the AI/ML development space. By recognizing community members as contributors to AI/ML development in authorship, researchers can work towards more equitable sharing of benefits.

2. *Alternative models of AI development*. In contrast to enriching technology companies' commercial foundation model offerings, some initiatives explore how to best support communities in developing their own smaller, more bespoke models, which are then owned and operated by community members. Past efforts have surfaced how communities often need to overcome *infrastructural barriers* such as limited available training data [53], capacity, and access to financial capital and compute [54] to support creating, hosting, and maintaining their own models.
   (a) One prominent example is the Te Hiku Media foundation, a Māori nonprofit, that decided to develop its own data hosting platform and transcription models for the *te reo* language [55].
   (b) Researchers from DAIR [56] have similarly urged the research community to support local indigenous NLP organizations like Ghana NLP and Lesan AI who "create machine translation systems for the specific communities they belong to".
3. *Alternative models of dataset ownership and usage.* Many participatory engagements involve compensating community members in exchange for complete ownership over their data (to use for future AI development). In contrast, several communities that own their data have experimented with alternative models to govern who can use their datasets or models, and for what purpose. These usage restrictions are often specified in *licenses* or other types of contractual agreements [57, 58].
   (a) Some licenses attempt to protect participants from AI-mediated harms by restricting how other stakeholders can use resulting datasets and models. The "Licensing African Datasets" project explores how to create licenses for African datasets that better redistribute benefits back towards African citizens and companies, with the expectation that "users in developed nations would perhaps pay for use of the work or use the work under more restrictive terms" [59]. Similarly, Te Hiku Media created a data license that "will only grant data access to organizations that agree to respect Māori values, stay within the bounds of consent, and pass on any benefits derived from use back to the Māori people" [55].
   (b) Future licenses can also explore specifying alternative compensation structures that allow communities to receive *continued royalties* (beyond one-time compensation) to encourage profit-sharing as models that depict their likeness continue to be used [60].
4. *Supporting community-driven impact assessment, criticism, and refusal.* Many participatory engagements motivate community members to participate by lauding the benefits of improved GenAI representations. We urge those conducting such engagements to *involve community members* in interrogating what barriers stand in the way of realizing these benefits, and in understanding potential algorithmic harms that may result from improved representations.
   (a) Facilitators of such engagements should make room for outcomes where participants decide the harms outweigh the benefits [61, 62]. For example, although queer scholars noticed that state-of-the-art AI voice cloning tools underperformed when cloning the voices of gay speakers, they ultimately decided against developing an improved AI technology due to concerns that an improved technology may be misused to surveil, misappropriate, or mock gay people [37]. Making room for such critical engagements will require educating participants who enter into engagements with varying levels of familiarity about AI capabilities and harms. We believe that facilitators similarly have much to learn from the situated expertise of community members – in fact, many scholars have argued that *impacted communities themselves* are best equipped to anticipate AI harms [63, 64, 65, 66].
   (b) Communities should not just be relegated to red-teaming roles where their cultural expertise is used to identify AI harms, as this can be psychologically damaging [67, 68] and further reify existing power distributions between AI developers and communities[2]. Rather, more work is needed to build the infrastructures that empower community members to define and achieve algorithmic accountability and recourse on their own terms.

[2]See the reports put out by the Wiezenbaum Institute's Data Workers Inquiry for more: https://data-workers.org/

For example, researchers can investigate how to support the translation of community-identified AI harms into implications for policy design to shape AI regulation following community needs.