Empowering Community Driven Determination of Values for Language Models

by

Deepika Raman

B.E., Bio-Medical Engineering, Anna University, 2017

Submitted to the Institute of Data, Systems, and Society
in partial fulfillment of the requirements for the degree of

MASTERS OF SCIENCE IN TECHNOLOGY AND POLICY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2024

Authored by: Deepika Raman
Institute of Data, Systems, and Society
May 20, 2024

Certified by: Dr. Dylan Hadfield-Menell
Assistant Professor of EECS, Thesis Supervisor

Accepted by: Dr. Frank R. Field III
Interim Director, Technology and Policy Program
Senior Research Engineer, Sociotechnical Systems Research Center

Empowering Community Driven Determination of Values for Language Models

by

Deepika Raman

Submitted to the Institute of Data, Systems, and Society

on May 20, 2024 in partial fulfillment of the requirements for the degree of

MASTERS OF SCIENCE IN TECHNOLOGY AND POLICY

ABSTRACT

Emerging technologies like Artificial Intelligence and Large Language Models are often developed in Western contexts and carry implicit values, from developer choices or underlying training data, which are not adequately representative of the diverse contexts in which they are deployed. The resultant misalignment from the lack of engagement with non-Eurocentric value paradigms results in inadequate, and potentially harmful outcomes that impact these unconsidered communities. To codify fundamentally subjective human values therefore necessitates the elicitation of these nuances through the inclusion and involvement of these very communities.

This thesis argues that participants' lack of familiarity with new technologies like Artificial Intelligence impacts their engagement and contribution to participatory processes of AI development. This thesis also helps demonstrate how grounded theory approaches can be leveraged to contextualize awareness-building efforts that can potentially empower community participation by addressing such familiarity gaps.

This two-fold objective of (i)eliciting community-relevant attributes for language model alignment (ii)through the necessary familiarization of the technology in question is demonstrated through the means of sample case studies. A grounded participatory process-CALMA (**C**ommunity-aligned **A**xes for **L**anguage **M**odel **A**lignment) is designed and evaluated through these cases to illustrate this contextualized alignment exercise. Learnings from this comparative case study are then extended to explore avenues for communities and institutions to adopt similar techniques that center the voices of the final users.

Thesis supervisor: Dr. Dylan Hadfield-Menell

Title: Assistant Professor of EECS

Acknowledgments

This research was made possible by the Algorithmic Alignment Group at MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL). I am extremely grateful to my supervisor, Prof. Dylan Hadfield-Menell for guiding me through this rare opportunity to work in such a multidisciplinary environment and also for bringing together such a rare crew at the lab whose genuine resolve to make tech work for *everyone* was the singular highlight of my academic experience at MIT. I am most grateful to have been on this research journey with Prajna, for the rewarding partnership, and for being infinitely kind and patient besides being wicked smart.

The most important contributors to this work were our participants, our collaborators who brought a deep curiosity to every session and generously shared their insights and reflections which enriched our user study. Balancing EST and IST time zones meant some very late nights and early mornings but every session only served to reiterate the motivation of this study to center communities. Our many interactions gave rise to some incredibly unique perspectives and all the hours spent rewinding and revisiting these conversations have only made me a better researcher. I would name all these incredible people if that wouldn't completely violate the IRB!

And, the biggest of thank yous are reserved for my lovely family and friends: To my parents, Kala and Raman, I thank you for listening to my endless chatter (work and existential) and for always embracing my moody enthusiasm, my confusions, my rants and my hopes; to my *paati*, Kaveri for being the light of my life; to all the women in my life Ishika, Anushka, Varna(x2), Mridula, Ilavarasi, Ramya, Mehr, Arundhati, Ranjini, Taylor and Sacha for the love, laughter, and courage.

Thanks finally to Rameen and Layla for the collective griping and solidarity through TPP, hand-holding me through the (unsuccessful) American acclimatization, and for creating a neurodivergent ~~haven~~ celebration at home every day. And Sia, you did the most of them all at Ellery.

My final thanks are owed to the SAGE community at Kresge that made it possible for me to keep going at a time like this and served as the only site of moral reflection and courage on an otherwise stifling and complicit campus. Your solidarity, or rather the collective struggle, provided the nourishment that sustained me through this final stretch- the kind of care only community can offer. Your dedication to liberation was an honor to witness and will serve as a lifelong inspiration.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Artificial Intelligence (AI) applications are increasingly becoming commonplace in today's world dictating many aspects of everyday life - from accessing customer support for e-commerce orders to determining eligibility for life-saving insurance plans. AI innovation has significantly improved the efficiency and dependability of several processes through automation[1], however, there is also ample documentation of the risks and harms that have accompanied these technologies. This has also been accompanied by the rapid erosion of end-user agencies from opting out of this algorithmic society. Proactively studying and mitigating these risks and their interaction with society has consequently been a critical focus in AI development.

Large Language Models (LLMs) in particular, have exploded into the emerging tech arena as a popular foundational technology on which a variety of applications are being built. These technologies have highlighted a unique set of harms[2]–[4] sometimes arising from the kinds of data they are trained on that, causing a steady accrual of deficiencies in these applications[5][6].

These harms could also be attributed to the development of such foundational models mostly restricted to resource-rich geographies of the West [7]–[9] but availed for a wide range of applications that are designed and deployed for other diverse contexts, largely reproducing homogeneous outcomes[10]. In these instances, the lack of necessary contextualization to incorporate relevant communities' preferences and values reinforces existing power structures and results in real-world repercussions[11] when the intended audience diverges from the

Western conception of a prototypical user. Especially owing to the inextricable role that language plays in shaping and being shaped by culture and community[12], artifacts like LLMs could have disparate impacts on marginalized sections of society when communities aren't centered in their design.

The Alignment Problem

Several technical methodologies such as supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) have investigated alignment strategies to encode values into AI systems [13][14] [15]. Often these are either mono-dimensional metrics of "human preference" or standardized researcher-prescribed values along which alignment happens [16]–[19]. While this has still helped make significant strides in optimizing algorithmic applications, it belies the notion that a "single reward function cannot represent a diverse society of humans."[20]

Key questions haunting AI alignment continue to be (i) *how do we define alignment, (ii) what are these values that we are aligning for? And, (iii) whose values are we aligning these algorithms to?*[21]

Any endeavor to identify the overarching value or set of values to guide all alignment therefore may not possess one right answer and would lead us to the same *formalization trap* [22] that burdens fairness literature. Most studies recognize the impossibility of arriving at a singular set of preferences and have instead attempted to restrict the alignment exercise to social preferences where there *is* consensus through partial order compliance[23].

Finally, the question of who the human in the alignment exercise is has been often addressed by the insufficient premise of "crowd-sourcing." As an emerging type of work in the gig economy, the data annotation industry is fueled by transient workers who may not sufficiently represent a multiplicity of cultures or contexts. Even when recruited to be a representative group, annotators are often treated as interchangeable identities indicating the failure of the task or platform design to successfully elicit their multiple perspectives[24]. Furthermore, they are assigned tasks within the prescriptive paradigms of researcher-prescribed (or more recently expert-designed) categories or predefined metrics. These categories range from sometimes reducing alignment to traits like *helpfulness, honesty, and harmlessness*;[25]

to include categories from standardized lists like the BIG-BENCH,[26] EvalPlus,[27] or the ETHICS[28] dataset containing scenarios about justice, deontology, virtue ethics, utilitarianism, and commonsense moral intuitions; TRUSTGPT[19] for evaluation of three ethical perspectives: *toxicity, bias, and value-alignment*, etc.

Such approaches result in the embedding of the developers' values into these applications[7] who often tend to be white, affluent, technically oriented, and male.[29] These contested concerns result in the continued absence of meaningful evaluation metrics against which to gauge LLM performance and pose critical accountability concerns in downstream use cases built atop these models.

"The Participatory Turn"

To address some of these shortcomings and the growing mistrust of AI systems, there has been a move to incorporate participatory design approaches to orchestrate a semblance of inclusion and equity to win back trust in these applications. Participatory design involves engaging diverse voices (end-users, stakeholders, and affected communities) in the design and deployment of algorithms, ensuring that the technology is shaped by those it affects. Participatory design and participatory action research are two disciplines that have widely shaped this area of inquiry, with successful adaptations often seen in the domain of Human-Computer Interface (HCI) research and design.[30] Many endeavors have helped showcase the rewards of these participatory approaches which have helped in advocating for better digital rights[31], identifying better design,[32] building governance frameworks, and exposing the underlying tensions between the values of the communities and the algorithms.[33] [34]

Emerging literature has also similarly shown an increasing attempt to take a more participatory approach to the alignment problem[34]–[37] to collect diverse preferences while accommodating the inherent subjectivity of value alignment tasks[38]–[40]. The demonstrated uniqueness of outputs from such studies serves as a promising sign on which to build the tenets of this thesis.

The Challenge

Building on this work to center communities in the participatory processes, a major challenge that emerges is **participants' lack of familiarity** with the technology or process thereby curtailing their ability to provide informed feedback on their individual preferences. This opens up avenues for the researchers' beliefs to seep in and influence participants' behaviors or interactions through priming or interpretation of results. Johnson and Verdicchio [41] show that a lack of clarity in the AI development pipeline results in a *sociotechnical blindness* that tends to underplay developer agency while bestowing anthropomorphic traits to AI systems, signaling their autonomy. Cugurullo [42] similarly critiques these simplistic notions that serve as "shortcuts" that distract from the reality of power systems at play. These incomplete perceptions of AI continue to limit participant voices that have so far been excluded from shaping these artifacts and minimize their potential to tangibly shift the power upheld/perpetuated by them[43]. Therefore operationalizing participatory processes for communities in diverse contexts cannot happen in the absence of an accompanying intent to empower users with the necessary knowledge of these technical systems.

While this attempt at community value alignment is necessary to negotiate power in today's inescapable pervasiveness of AI systems, it is also at odds with the inherent subjectivity of interpreting human values[44]. To reconcile this contention and allow for multiple notions of values to emerge unbridled by prescriptive methods, it would be necessary to adopt a grounded theory approach toward facilitating these outcomes. Such a strategy not only helps to minimize researcher bias but also enhances participant agency in shaping both the design and outcomes of these participatory processes.

This thesis argues that participants' lack of familiarity with new technologies like Artificial Intelligence impacts their engagement and contribution to participatory processes of AI development. This thesis demonstrates how grounded theory approaches can be leveraged to contextualize awareness-building efforts that can potentially empower community participation by addressing such familiarity gaps. The main areas of inquiry towards this will include (i) a discussion on the tenets of value alignment and participatory design, best practices, and challenges from liter-

ature, (ii) a user study of a grounded participatory process built on this literature to elicit community-relevant attributes from two distinct groups of participants, (iii) a qualitative analysis of the study's familiarity building exercise, and (iv) articulating considerations to promote better incentivization of similar participatory approaches through community and policy levers.

Chapter 2

Capturing Nuance and Facilitating Participation

This chapter examines literature that demonstrates AI harms from misalignment and explores current language model alignment strategies. The recent explosion of participatory techniques to reduce AI harms and aid alignment is also surveyed along with critical inputs on their design. Importantly the articulation gap in these participatory methods arising from the recency of these technologies and participant unfamiliarity is also documented. These insights help underscore the significance of context-specific and community-informed evaluations.

2.1 AI Impact on Communities

Non-Western Narratives:

Investigating harms from AI systems and language model applications are often restricted to identities of race and gender but many studies define and expose harms perpetrated on members of other marginalized communities by these technologies. Through focus groups with participants with disabilities. Gadiraju, Kane, Dev, *et al.* [45] identify stereotypes and knowledge gaps about disability from language models response and present the need for developers to co-design annotation and training processes with people with disabilities.

Khandelwal, Tonneau, Bean, *et al.* [4] demonstrate poorer language model performances when evaluated for non-Western biases through the example of an Indian context focusing on the identities of *caste* and *religion* as the axes and present mitigation strategies through instruction prompting. Through more computational approaches, Diaz, Johnson, Lazar, *et al.* [46] show how age-related biases manifest in popular ML-based sentiment analysis methods and suggest a custom classifier as one way to mitigate this. All papers ultimately allude to the larger reality of how technology, especially language technology, is inherently capable of affecting how power manifests in society, and present such community-specific investigation as an urgent need in providing insights on the rate of spread of such societal biases and the need to identify, document and mitigate them by centering the affected communities in question.

2.2 Technical Methods for Alignment

2.2.1 General Alignment

Recent literature has explored several technical methods for language model alignment which can broadly be categorized into RLHF-based methods or SFT-based methods [13]. RLHF involves fine-tuning a pre-trained model using a reward model that is trained on human preference comparisons between different outputs for the same prompt. There have been many variants of RLHF that have been explored in literature. [47] explores how rule-conditional reward modeling where preferences are broken down into rules in natural language can be incorporated into RLHF. [48] proposes SENSEI, a RLHF-based method that embeds human value judgments into each step of language generation through an Actor-Critic framework where the Critic simulates the reward assignment procedure of humans and the Actor guides the generation towards the maximum reward direction. There are also other variants that are RL-based where the model learns via text edits [49] and synthetic feedback [15]. Other recent technical alignment methods include Kahneman-Tversky Optimization (KTO) and Binary Classifier Optimization [14], [50] which aim to align LLMs using binary human feedback signals on prompt-completion pairs.

2.2.2 Value-based Alignment

There is also a growing subset of technical alignment literature focusing on value-based alignment. [51] proposes PALMS, an iterative process to significantly change model behavior by crafting and fine-tuning a dataset that reflects a predetermined set of target values. [52] propose Constitutional AI which uses RLAIF (Reinforcement Learning from AI Feedback) with a set of rules or principles referred to as the 'Constitution' for human oversight.

2.2.3 Multiple Dimension Alignment

To expand language model alignment beyond preferences along a uni-dimensional axis, works like SteerLM, FUDGE, and PPLM explore alignment along multiple attributes. SteerLM proposes an alternative to RLHF by using attribute-conditioned supervised fine-tuning to align LLMs with human preferences [53]. It conditions responses to conform to an explicitly defined multi-dimensional set of attributes like helpfulness, humor, and toxicity, enabling user control over the generated outputs. FUDGE (Future Discriminators for Generation) explores attribute-grounded generation which uses descriptions of desired attributes as conditioning to control text generation [54]. Directional Preference Alignment with Multi-Objective Rewards proposes a framework to align language models with diverse user preferences along multiple objective dimensions simultaneously, enabling fine-grained control over generated outputs [55]. These technical methods lay the groundwork for context-specific processes to build on.

2.3 Participatory Methods in AI

Significant benefits through the inclusive and responsible development of AI have been reported by using participatory approaches at different stages of the AI pipeline. Participation is often centered as a key driver in both platforming the negative impacts of AI on marginalized identities and empowering their agency to enact change[56]. Delgado, Yang, Madaio, *et al.* [57] document user-centered design, Participatory Design(PD)Participatory action research (PAR), Social choice theory (SCT), Civic Participation, Value-sensitive Design (VSD), etc. from design literature to supplement current approaches to engaging various

stakeholders of the AI pipeline and a framework to evaluate the recent "participatory turn."

Asaro [58] situates Participatory Design (PD) relationships in the historical methodologies that helped shape it, despite their being conceptualized for different relationships of the worker with the technology. While one movement aimed to protect workers and rectify the imbalances brought forth by technology in the workplace, the other aimed to increase the efficiency of these new technology systems. These inherent tensions in the premise of these two movements and their rare convergence to form the discipline of PD as it is understood today prompts the imagination of participatory approaches as not just a "design fix"[59] but as a critical process for the development of technology itself[58],

One shortcoming of current approaches to building AI products is their heavy reliance on modes of participation that attempt to include 'domain experts' in consultative capacities to opine on the labels/products/models through transactional engagements.[57] In these consultative setups for refining algorithmic accuracy and fairness, any derived knowledge is ultimately in the service of industry, not users. Additionally, the instances of participation are often in earlier steps of the AI value chain, at the level of data curation, annotation, and labeling - seldom at the model development and weighting phase. Often these modes of participation take the form of workshops, awareness campaigns, or other similar interactions that limit the transformative power of participation. While it has been crucial to involve more voices and perspectives in the development of AI tools, concerns have been raised regarding the turn to 'participation-washing'[59] where participant voices are selectively curated to resonate with predetermined aims or simply do not engage participants in a meaningful way in the process.

These cases also serve to dilute the responsibility of developers in the product cycle, while not transferring any ownership of the product itself. This choice attempts to obfusciate the question of accountability aggravating the 'problem of many hands' in the multi-stage ML pipeline.[60] Concurrent with this concern for accountability, the risk of rent-seeking behavior of products that co-opt participatory methods to purchase user trust cannot be dismissed.

2.3.1 A Side-note on Participation in AI

It is important to emphasize that all AI applications are inherently participatory regardless of an accompanying 'Participatory AI' tag[61]. Participatory work features in the perennially undervalued and underpaid data annotation industry, the (often undisclosed) labor of all the users interacting with these applications (whose interactions further train these models), and the contribution of digital artifacts (mined from monitoring activity/scraping web data)[62]. This unacknowledged web of participation largely extracts from underpaid labor in low and middle-income countries creating the rich corpora of data compared to (arguably) overpaid Silicon Valley developers for their contribution to the AI development cycle. This is illustrative of the dual risk of strengthening property rights while evading liability as shown by NissenbaumNissenbaum [63] for the computer industry, which has unsurprisingly permeated into the AI industry.

While these colonial[64] underpinnings pervade global AI development and need a re-imagination of AI development pipelines, a first step would require us to explicitly articulate the dynamics of power in these "Participatory AI" interventions and transparently register the scope of participation to shape the studies in question.

2.3.2 A Model Example of Participation

The strongest manifestation of participatory approaches in AI is best illustrated by the African languages machine translation project by Nekoto, Marivate, Matsila, *et al.* [65]. The exemplary self-organized network comprised 400 participants, from over 20 different countries including content creators in local languages, language translators, and experts in language technologies who contributed to the development of machine translation systems for underrepresented languages in Africa. By involving stakeholders in the data collection and annotation process, the Masakhane community was able to emphasize and document local contexts and linguistic nuances to produce truthful machine translations. This study is especially noteworthy in how it afforded absolute flexibility for participants to define the goals for the study, transition between roles in the community, experiment in different forms of engagement and ultimately ascribe a sense of control over the community's outputs

to the participants. The range of outputs from this study includes datasets, models, and benchmarks (46 benchmarks for translation models from English into 39 distinct African languages) available freely to the public. Masakhane continues to actively serve the translation ecosystem to address language disparities, contribute to the democratization of technology, and showcase a model for the institution of diverse teams in the formative stages of tech development to deliver trustworthy technical artifacts for their communities.

2.3.3 Meaningful Participation:

Frameworks[66] [56] to determine the meaningful participation of stakeholders in AI include metrics on the representation of stakeholders, their stage and duration of involvement in the ML cycle, feedback channels, participant empowerment, study setup, etc. as the relevant axes against which the effectiveness of such interventions can be evaluated. (i)Co-designing processes, (ii)accommodating plurality, and (iii)iterating collectively to build participant investment in the technology emerge as a repeating theme in the reviewed literature.

Designing *With*

The field of Participatory Action Research (PAR) could be leveraged here to enable setups that collaborate and 'design with' communities instead of 'designing for' them. While PAR methods may not be straightforward or seem practical to execute within the extractive nature of ML systems, the case study Feminicide[1] Counterdata Collection from the Data+Feminism lab[67] serves as another illustrative example of its success (and shortcomings). This initiative developed an ML model that was trained to retrieve relevant media reports on instances of feminicide (which typical search queries were inefficient with) to support the work of feminicide data activists. Unlike posthoc fixes that typically predominate the endeavor of fairer outcomes, this proactive approach allowed for unorthodox data collection and annotation initiatives, partnerships (as opposed to consultations) with activists as experts, and an intersectional feminist approach that guided the design and evaluation of the entire process.

[1]Feminicide is defined as the gender-related killing of women. See Camilo Bernal Sarmiento, Miguel Lorente Acosta, Françoise Roth, and Margarita Zambrano. 2014. Latin American model protocol for the investigation of gender-related killings of women (femicide/feminicide). New York: United Nations High Commissioner for Human Rights (OHCHR) and UN Women (2014).

It elucidates the leadership of activists in multiple stages of the technical development of the ML tool through knowledge-sharing of context and emphasis on intersectional identities in a collaborative, iterative model training exercise.

Encoding Intersectionality and Centering Context

When conditions for meaningful participation are met, there could be remarkable implications for both AI applications and their governance. The devolution model places people at multiple levels of the AI development process, providing them access to determining goals, parameters, weights, and extensions of the model. Participatory methods inherently deal with consensus building between the representatives of different groups and individual opinions. The requirement to handle competing inputs, centering algorithmic justice over statistical fairness, encoding intersectionalities, and translating particular contexts motivates the creation of more robust models.

Iterative Design

In the case of the Feminicide Counterdata Collection, the iterative nature of the case allowed for the collection of context-specific negatives that were discovered to be underspecified in the model. Such a scenario is generalizable to similarly complex problems with particular combinations of parameters leading to better outcomes. Similarly in the WeBuildAI[68] study, the process of individual model building to collective aggregation with participants positively impacted procedural fairness and distributive outcomes wherein participants believed that having control over the design of the algorithm made the results more fair, generating trust in the technology.

2.4 Participatory Methods for AI Alignment

2.4.1 Voting on Preferences

Prabhakaran, Davani, and Diaz [38] illustrate how label aggregation methods flatten human preferences and prescribe an annotator-level label retention to accommodate the subjective

nature of human preferences across socio-cultural differences which can be used to inform modeling steps involved in the downstream deployment of model-based applications. Recent literature has seen an increase in the use of participatory methods to enable such aggregations of preferences for alignment without minimizing multiple perspectives. Existing participatory approaches like Wikibench and PRISM aim to elicit community perspectives for AI evaluation and alignment. Wikibench enables communities to collaboratively curate AI evaluation datasets while navigating consensus, disagreement, and uncertainty through discussion [37]. PRISM maps the sociodemographics and stated preferences of 1,500 diverse participants to their contextual feedback on 21 LLMs, demonstrating how different humans can set divergent alignment norms [34].

2.4.2 Voting on Values

This category of alignment work has gone on to open up the very values along which preferences were elicited in the previous section. The "Veil of ignorance" approach proposes an alternative framework grounded in social choice theory to derive impartial principles for AI governance [69]. Participatory methods in this type of alignment that relate most closely to CALMA include Collective Constitutional AI and STELA. Collective Constitutional AI by Anthropic drafted a constitution with over 1,000 Americans by allowing participants to vote on statements relating to AI on Polis, a real-time interface augmented by statistics and machine learning to gather and understand what large groups of people think in their own words [36]. STELA proposes a method that applies participatory techniques to elicit rules for agent alignment through community-expert-defined themes and community-participant by assessing dialogue samples, these dialog samples are researcher-curated and final rule sets are researcher-derived by two of the authors [35].

2.5 Participant Perceptions of AI

2.5.1 Survey Findings on AI Awareness

Survey-based studies that sampled representative populations to understand public perception of AI include, Zhang and Dafoe [70] who revealed the mixed support that AI development received from the 2000 American surveyees. Cave, Coughlan, and Dihal [71] similarly document attitudes in the UK where only 45& of all respondents could describe what AI was as opposed to the 85& who claimed awareness of its existence. The survey also unearthed an anxious outlook in the UK public that was brought about by the suspected lack of agency that society or citizens reported in shaping the future of AI development. Balaram, Greenham, and Leonard [72] previously surveyed public familiarity with AI systems and registered a greater awareness of the more tactile or interactive pieces of emerging technologies like autonomous vehicles (84&) or digital assistants like Siri or Alexa (80&) as opposed to much lower recognition of chatbots (46&) and an even lower (32&) knowledge of autonomous decision-making entities. Even ones being deployed for making decisions in highly contentious applications like criminal justice, immigration, healthcare, etc. This was similarly corroborated by Ada Lovelace Institute [73].

Bewersdorff, Zhai, Roberts, *et al.* [74] review the literature on common misunderstandings of AI, especially in the educational context, and shine a light on the largely binary, unspecific, and often incorrect views that learners have about AI concepts and development that lead to both anxieties and misconceptions of unemployment from automation. They also point to a marked lack of focus on literacy efforts that focus on inclusiveness, bias, or trust which they flag as an unattended area.

2.5.2 AI Perceptions Across Cultures and Identities

Most studies investigating the public perception of AI have unsurprisingly featured Western audiences and present the need for a more cross-cultural comparisonCave, Craig, Dihal, *et al.* [75]. Surveys have also recorded perceptions of younger and more affluent populations as being more familiar with and favorable toward automated decision-making, conjecturing

that these groups expect technological advances to be beneficial to them[72]. Apart from socioeconomic status, significant correlations between user identities (of race, gender, etc.) and their corresponding opinions about AI have also emerged[76]. Similar findings emerge on the AI awareness and trust reported by male students compared to female students[77] as well as a recent survey by Brian Kennedy and Saks [78] indicating higher AI awareness as correlated to higher education and a similar gendered divergence.

This pattern is not unique to AI perceptions and has been true of tech reception of the past as well, with varying levels of trust reported across identities attributed to a combination of cultural and historical legacies[79]–[82].

2.5.3 Prior Work Enabling AI Familiarity

The field of Explainable AI is central to addressing the "black-box" problem of AI systems and presented several methods to tackle transparency both through explanations for human understanding at different stages of development for different types of stakeholders[83]. While these approaches provide the necessary transparency and interpretability of these systems[84] there is a more urgent need for AI literacy to equip end-users with the necessary competencies[85].

Several studies have attempted to address tech familiarity through novel ways, Kihara, Bendor, and Lomas [86] for instance deploy a creative "Escape the Smart City" game to build resident familiarity with the complex sociotechnical ramifications of the use of facial recognition technologies for state surveillance in the city of Amsterdam. Robb, Ahmad, Tiseo, *et al.* [87] leverage social psychology models of Scientific Literacy to enable participants to make informed decisions through interactive activities that involved controlling different robotic exhibits alongside survey quizzes that recorded participant knowledge with no significant impact.

Richardson, Prioleau, Alikhademi, *et al.* [76] methodology involved eliciting participant trust (defined as 'willingness to use') and perceived benefit to their community (defined as their hometown or family) by introducing them to the field of AI and its application (in healthcare, law enforcement, etc.). The paper however does not provide any further detail on this familiarity exercise except reported participant satisfaction with such an awareness

exercise.

2.5.4 Lessons from Other Tech Awareness Efforts

Ballard, Werner, and Priyadarshini [88] document the limitations of language in adequately enabling the accurate understanding of system dynamics to enable participatory modeling and consensus building in cross-cultural and/or multi-lingual settings. Through the case study of a visual mapping exercise conducted in rural India, they demonstrate participant empowerment to negotiate nuances of different variables and reinterpret the socio-ecological interactions resulting in a first-of-its-kind level of ownership and engagement from a large set of stakeholders in modeling for their diverse multi-lingual community.

D'Ignazio [89] steers clear of the strict definitions of literacy and its perception here as purely technical knowledge and instead proposes tactics to empower a non-technical audience through the creating data biographies that tell the story of data collection and organization, building learner-centered tools all premised on immersive community-centered themes[90]. The Internet Democracy Project's embodied approach to data takes on a unique spin in illustrating the fading distinction between physical and virtual bodies and presents any exploitation of virtual data, akin to a violation of bodily integrity to drive home a rights-based approach to empower participants' conception of the consent process.

To borrow from university pedagogies, three tactics emerge as key drivers adapted to facilitate a range of learning goals for AI namely, visualization, real-world contextualization, and domain specificity[91].

2.5.5 Situated Knowledge for Participatory AI

The Algorithmic Equity Toolkit[92] presents a flowchart method for community organizers to identify AI use in state surveillance and its potential for harm, empowering leaders to defy surveillance by supporting necessary political action to create social change. This method was not without its limitations wherein different stakeholders in attendance like-community organizers and data scientists, had differing understandings of the computational concepts under study, causing divergent interpretations.

In enacting sociolinguistic justice into action, a student-centered curriculum that celebrates linguistic diversity while recognizing linguistic racism centering high school youth as linguistic experts through the SKILLS Program for linguistics[93].

2.5.6 The Cost of Poor Articulation

In failing to adequately address the familiarity gap in participatory techniques, especially involving high-stakes algorithmic content, any PD techniques risk misattributing resultant outcomes as community-driven insight when it is ripe with misinterpretation, or with researcher bias. The cascade effect that threatens to follow, in terms of scaling technical output or furthering policy aims, could actively harm the (often) marginalized communities under consideration. In the undeniable premise of communities existing within inextricable structures of power, the cost of inadequate context setting for participants, any outcome, especially in Global Minority settings, should be treated as an extending legacy of colonial extraction.[88]

Chapter 3

Methodology

While positivist epistemologies have almost exclusively supported scientific inquiry, they fundamentally conflict with the inherently interpretivist nature of human values. This thesis is an exploratory study aiming to identify a participatory process that adequately accounts for individual subjectivity while facilitating collective consensus as we attempt to align AI with values that communities find relevant. Much like the ethos of the CALMA process piloted in this study, this thesis also takes a hypothesis-free approach to re-imagine language model alignment and evaluation.

Participatory Design:

Participatory Design (PD) and Action Research (AR) are two collaborative approaches that are often leveraged in HCI research despite their contested generalizability. However, it is precisely because of the ability to develop localized solutions that motivated the choice of this methodology for this community-centric study. Generalizability while not the goal of the value-alignment exercise can instead be replaced with the cross-contextual transfer of participatory processes (here CALMA) to achieve other hyper-specific results that are arguably more relevant to the goals of this study.

Through its historic roots, Participatory Design is fundamentally concerned with values, and the choice to adopt such a methodology inherently reflects the values that participants and researchers bring into such a process[94] . Rorty [95] presents an edifying philosophy that anchors Action research as an ongoing dialog and debate, geared towards achieving

"communicative clarity," not a ground truth. This methodology is adopted for this inherently subjective value alignment exercise owing to its explicitly non-hierarchical nature that enables creation "with" people instead of "for" them through iterative, co-designed processes.

While PD emphasizes design and AR emphasizes research, the ethos of both methods rests in the intention to solve sociotechnical problems collaboratively. PD largely determined the first phase of the study while AR techniques were borrowed to co-create and deploy Phase Two. Creating an environment for dialog and discussion in both phases of the study would therefore result in an easy articulation of values grounded in data[96].

Emergent Design:

This thesis uses inductive reasoning to lend structure to the patterns that emerge from designing and evaluating the CALMA process in a context-sensitive way. The emergent perspective adapted for this required the denouncement of rigid research flows that do not accommodate an adaptive design as the understanding of the area of inquiry expands. Often emergent methodologies are best suited for studies that are carried out over a longer period whereas shorter studies tend to follow a standard model. However, this is often premised on the idea that research questions have sufficient answers from a singular study as opposed to a collective model of knowledge production in the field [97]. While the CALMA process is not entirely emergent in its design, the ability to co-design the second phase of the study through participant interviews was actualized only because of this methodological choice. Design inputs from participants presented us with unanticipated opportunities to glean further nuance of this value labeling exercise, indicating the mutual benefits of such a methodology[98]. At this juncture, it bears mentioning that such an adaptive design is fundamentally at odds with the extensive pre-planning required for protocols involving human subjects research that need approval from Internal Review Boards (IRB).

However, the motivation for the choice of an emergent design was not philosophical alone, but also in response to the practical contingencies that arise from coordinating a study with 26 individuals across time zones, where their investment in the study cannot be assumed through their consent to participate alone. Seeking access to participants' time is, and should be, a continuous negotiation and only a flexible research plan can accommodate these

practical considerations to prioritize participant agency in these forums. While extensive scheduling considerations went into each session of this study, the final group discussion best illustrated the need for contingency planning wherein significant participant attrition was observed between phases one and two. The flexible design of the CALMA process was, therefore, necessary to accommodate this drop in numbers, which presented a unique set of observations that enriched the findings in this study.

Grounded theory:

The study design was preceded by a comprehensive literature review to understand the gaps in current participatory techniques. One repeating theme can be illustrated through Bergman, Marchal, Mellor, *et al.* [35] 's participatory method that elicits rules for agent alignment by making community participants assess researcher-curated dialogue samples. The consequent community assessments of researcher-prescribed adversarial chats in this study, even when informed by expert-led topics, tend to produce a set of rules that cannot necessarily be considered free of researcher bias. This prompted the choice of a grounded methodology as the preferred choice for this study to construct theories (derive community-relevant axes) in a non-prescriptive manner, grounded in empirical data (model interactions).

In the context of Language Model Alignment, Charmaz [99]'s Grounded Theory approach was chosen to help uncover the implicit and explicit values, beliefs, and norms that language models exhibit by analyzing language model responses. As with the tenets of grounded theory, this study leveraged an open-ended annotation process, through two stages- initial coding and selective coding, to avoid any influence of standardized evaluation parameters or value lists from unduly influencing participant labels.

More specifically, this thesis adopts Charmaz's Constructivist GT[100] which espouses a form of reflexivity that prompts a continuous questioning of the process of inquiry of data collection and analysis. This was facilitated through the presence of two distinct participant groups between whom different iterations of the inquiry process could be tested. This category of Grounded theory was also chosen because of the inherently subjective notion of such a value association exercise of data annotation. To accommodate these subjectivities instead of seeking ground-truth labels, the constructivist approach fits best because of its

underlying assumes that "what we take as real, as objective knowledge and truth, is based upon our perspective . . . "[101]

Studies leveraging this methodology often begin with purposive sampling, which in this case was the recruitment of two sets of participants to pilot the case studies. The research design then followed concurrent data collection and analysis through coding techniques (via participants interacting with the chatbot and annotating values identified), followed by a comparative analysis (through affinity mapping to create the embedded space of all artifacts generated through coding), and theoretical sampling (here through the dialogic inquiry enabled by a group discussion to register different perspectives, expand definitions, elicit further chatbot examples, etc.), which is done iteratively until theory saturation (arriving at community-relevant axes).[102]

Affinity Mapping

Affinity mapping is a common method from HCI often used to analyze data by categorizing it according to themes or ideas to extract insights and identify patterns[103]. It is especially used in team setups to detract from any creep of hierarchy in the data or ideas under review and enables a collaborative process of research organization and analysis. This technique is often enriched by an iterative mapping process to enable a more exhaustive pattern recognition exercise. Similarly using contextual information and other metadata also enriches this process[104].

In order to minimize researcher bias in the analysis of labels and annotations produced by participants, affinity diagramming was used to identify similar attributes and place them in neighboring clusters in an embedded space for Phase Two of the CALMA process. This was done iteratively by the co-authors to identify similarities both from the semantics of label names and also from the contextual information of participants' understanding of the focused coding exercise. This embedded space was mapped out through several rounds of discussion to minimize disagreement.

Affinity mapping is often sidelined as an inherently subjective exercise that limits the generalizability of the analysis. This is precisely why this method was best suited for this study. With the limited sample size, affinity mapping was an ideal tool to create a visual

representation of the value spectrum for participants to dissect and discuss.

Semi Structured Interviews

Synchronous online interviews were conducted one-on-one with each participant after their completion of each phase of the user study in Chapter 4. A semi-structured approach was taken to accommodate the registration of participants' perspectives in a flow they felt most comfortable with. Owing to the interpretivist approach of not just the analysis of gathered interview data but also because of the contents of these interviews dissecting yet another interpretivist task of value association, there was an attempt to build better rapport with each participant. The interviews served as an extension of the open-ended nature of the study allowing participants to register their experiences and feedback. These interviews were not free of the researcher's own values and interview philosophy, causing it to often be dialogic. This served to increase participant familiarity with the technical artifacts in question and in facilitating the action of knowledge sharing to benefit the participants as well. All interviews were recorded for transcribing purposes after receiving due consent.Qualitative data analysis software NVivo was used to code interview transcripts[105] and followed by stages of grounded coding to identify key findings.

Chapter 4

Deriving Context-Specific Axes for Language Model Evaluations and Alignment

Note: This chapter was co-authored with Prajna Soni at the Algorithmic Alignment Group at MIT CSAIL as part of a conference paper submission.

4.1 Methods

4.1.1 Process Design

Emerging literature has shown an increasing accommodation for the inherent subjectivity of value alignment tasks [38]–[40], and participatory methods [34], [37] to collect such value preferences from communities for alignment tasks. Building on this work to center communities in the alignment pipeline, some challenges from participatory methods that are further investigated include:

1. **researcher bias**, where the researchers' beliefs influence the participant's behaviors or interactions through priming or interpretation of results

2. **lack of familiarity**, where participants have not used a technology or process enough

or know enough about how a technology works to be able to comprehensively provide feedback on their individual preferences

3. **complexity and scalability**, where engaging a large number of participants from diverse backgrounds poses a logistical and methodological challenge

With these challenges in mind, CALMA (**C**ommunity-aligned **A**xes for **L**anguage **M**odel **A**lignment), a non-prescriptive and grounded process was designed to elicit context-specific axes while minimizing researcher bias. CALMA involves 4 key elements, each of which helps address the challenges of participatory design. At each step of Figure 4.1, we aimed to ensure a participant-centered approach by designing a non-prescriptive and open-ended process. Below, we describe the methodological choices of CALMA.

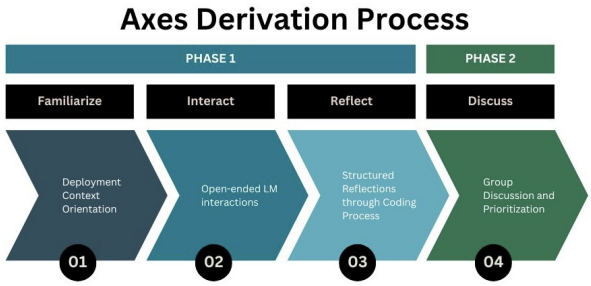

Figure 4.1: The CALMA process. Our study took place over two phases with four key elements: (1) Contextualizing the Deployment, (2) Open-ended LM Interactions, (3) Reflecting via Open Coding, and (4) Group Discussion and Prioritization

[1] Familiarize: Deployment Context Orientation

The first step of the CALMA process builds participant familiarity with the CALMA process and the deployment context of the language model. We presented a pre-recorded video that visually outlined this alongside examples of Phase 1 grounded in an unrelated deployment context to minimize any priming before they engaged with the model. An open challenge in participatory processes in the context of emerging technologies like LLMs is participant unfamiliarity with the technology. The Familiarize step aims to mitigate any impacts of this on the process outputs.

[2] Interact: Open-ended LM Interactions

Users were asked to approach the prompting exercise (1) in the context of the language model's deployment and (2) identify any implicit and explicit values that the model exhibited. The instructions provided were meant to minimally prime users and not place any restriction on the kinds of topics they chose to prompt the model. The language model interface for this step was built using Llama-70B and Mixtral-8x7B where a system prompt was appended to the interactions to contextualize the model responses.

[3] Reflect: Structured Reflections through Open Coding

The structured reflections were rooted in Charmaz's Grounded Theory Coding which leverages an inductive technique for gathering and analyzing data iteratively. We adopted a line-by-line representation of the interaction data from Step 2 for participants to code in two stages by modifying Price [106] OpenCodingForMachineLearning tool[1]

1. **Initial Coding**: The set of interactions from the first segment was exported to the tool where participants began annotating the text with values/qualities they observed implicitly or explicitly. The set of associated values/qualities identified by users in this segment are hereon referred to as "**annotations**". The open-ended nature of grounded theory allowed participants to assign one or more potential values they identified in the model's behavior.

2. **Focused Coding**: The dataset of interactions and associated annotations were then presented in the focused coding interface. The objective was to identify annotation clusters and assign them to groups that better identified them collectively. The group names from this stage will be referred to as "**labels**".

[4] Discuss: Group Discussion and Prioritization

Following the initial and focused coding sessions, users discussed their labels with other participants from the group in an open-ended format to arrive at a set of attributes with

[1]Source code for the tool is available here

definitions and examples ranked as top 3 and top 5. The goal of the discussion session was for participants to build consensus and produce a set of "relevant" attributes for their communities. In keeping with the non-prescriptive format, the design of this session was informed by inputs from participant interviews at the end of Phase One. The discussion was supported by a Miro Board[2] allowing participants to collaboratively engage on a virtual platform.

1. **Session Artifacts**: Participants were presented with three artifacts before the scheduled discussion, (i) their interactions, annotations, and labels, (ii) a summary document with topics the group explored and a word cloud capturing the frequency of the groups' annotations based on a simple word count, and (iii) an optional video that contextualized the outputs of the study in the LM alignment pipeline.

2. **Session Orientation:** Participants were introduced to the objectives of the discussion. The guidance for this session was limited to the following key pointers:

   (a) *There are no wrong answers*: the exercise's focus is on capturing subjective perspectives, not ground truths.

   (b) *Frequency does not necessarily indicate importance*: annotations that are uncommon could still be relevant to your community.

   (c) *Dialog not debate*: diverse perspectives must shape this conversation but the goal is still collective consensus

   (d) *It's okay to change your mind*: annotations and labels from Phase One are here to guide us and not contain us.

3. **Segment One: Initial Individual Ranking of Attributes**: At the start of the group discussion, each participant submitted their top three attributes with definitions after reviewing their interactions from Phase One. The goal of this was to allow participants to independently think about values and qualities that emerged from their interactions and form their preferences before engaging with others' arguments in the discussion.

[2]Miro is a digital collaboration platform designed to facilitate remote and distributed team communication and project management.

4. **Segment Two: Exploring Embedding Space**: Participants were asked to a visual embedding space on the Miro interface. The embedding space was created by the researchers which arranged labels based on their semantic meaning alongside all of the annotations that were grouped under each label. For example, 'balanced', 'diplomatic', and 'nuanced' would be close in the embedding space whereas 'biased' and 'one-sided' were clustered together. To mitigate researcher bias, clustering was restricted to the semantic similarity of the labels without any additional insights from the interactions they were associated with. The intent of the visualized embedding space was to condense the collective interaction data of the group into a more understandable format. (See Appendix for a snapshot of the embedded space)

5. **Segment Three: Individual Presentations**: Each participant presented their top three attributes to the group, explaining their interpretation of that value as observed in their interactions with the model. This provided an opportunity for everyone to contribute and break the ice. Finally, each attribute with its accompanying definition was added to the Miro board for others to review and reference throughout the discussion session.

6. **Segment Four: Group Discussion**: Participants engaged in discussion to understand each other's perceptions of the task at hand, the values they individually coded, any reactions or clarifications they had to attributes on the board, and continued dialog to arrive at axes that were relevant to their community. The final list was supplemented with the group's definition of the attribute and examples to ground it in model interactions. The group then **collectively ranked** these values as their top three and top five.

7. **Segment Five: Final Individual Ranking of Attributes**: After the discussion and final ranking of collective attributes and their definitions, each participant individually rated their agreement with the group's attribute definitions on a five-point Likert scale. Additionally, they submitted their individual top three and top five ranking of attributes they found pertinent following the group discussion. Segments One and Five helped illustrate how the discussion contributed to individual perceptions of the

attributes and their role in consensus building.

4.1.2 Evaluation Methodology

To evaluate the process design, we included surveys and interviews with participants following each phase to obtain feedback and identify opportunities for improvement through co-creation.

Phase 1:

1. **User Survey**: Following the Familiarize, Interact, and Reflect process, users filled out a survey set to a five-point Likert Scale rating across 15 prompts capturing their impression of the tool's interface and the task at hand. The survey ends with an open-ended question to gather any other inputs they have on the tool and process. This allowed us to understand the participants' attitudes to the technical artifacts supporting the coding process and the order of the tasks from Phase One.

2. **User Interview**: Participant engagement with Phase One culminated in a semi-structured one-on-one interview to gather their impressions of the grounded coding session. Participants were also briefly informed of the structure and intention of Phase Two and were invited to share any inputs or concerns they had.

Phase 2:

1. **User Interviews**: One-on-one interviews with participants were conducted to record their reactions and reflections on the discussion. Specifically, we wanted to gather feedback on their experience in an assessment of the consensus-building exercise and understand how it influenced their perceptions of the attributes by inviting them to reflect on their initial versus final ranking of the attributes.

4.2 Case Study Approach

To evaluate the process and understand how such a process would fare within different communities, we recruited participants from two distinct groups to participate in an IRB-approved study to derive community-relevant axes for a “History Educational Assistant" for high school students. History was chosen as the subject as it varies contextually and is subjective to a large extent.

The first group was a pilot group consisting of 11 graduate and undergraduate students from MIT who were US citizens, and the second in-context group had 15 working professionals from India. Both sets of participants were comfortable with English as a conversational language and had at least an introductory exposure to language models (e.g., were aware of their existence and could name examples of LMs and potential use cases).

Each participant was compensated at a rate of USD 17 per hour for the time they spent across both phases of the study. All sessions for Phase 1 (Familiarize, Interact, Reflect) were conducted online over Zoom. Phase 2 was conducted in person for the MIT student group and over Zoom for the Indian participants.

Since the premise of this study was to test the grounded and non-prescriptive CALMA process, the groups were not sampled to be representative or a set of experts representing a given community, so the final attributes cannot be ascribed as indicative of any community's preferences.

4.3 Evaluating The Process

In this section, we evaluate CALMA through the data and observations we gathered from running this study with two different populations and share insights on how design decisions impacted the output and effectiveness of the process.

4.3.1 Training / Educating Participants

From our case study sessions, we had two primary takeaways that could minimize the multiple dimensions of variance observed from task ambiguity:

1. engaging in such a specialized and time-intensive process should be preceded by necessary training along with an element of testing to ensure a baseline level of understanding of the tasks, and

2. Furthermore, increasing participant familiarity by contextualizing the study tasks into broader applications and tangible use cases is necessary.

Train and Test

Despite the intention of the Familiarize segment of Phase One, and the audio-visual orientation that accompanied the intention and objectives of CALMA, its interface, and the coding tasks, participants did not always grasp the grounded coding process. This indicates the need for **more comprehensive coding training** and **including an element of testing to ensure a baseline level of understanding of the task** at hand.

In follow-up interviews, a few participants expressed a preference for a preset word list they could build on to support their initial coding exercise. They attributed the difficulty in producing unique labels from scratch to (i) unfamiliarity in recognizing values in LM interactions, and (ii) the limitations of their vocabularies in generating more appropriate/detailed annotations. All information, instructions, and examples from the Familiarize stage were also available to the participants via a concise document for participants to refer to throughout the Reflect stage. While the goal of the study was to identify 'values and qualities' in their interactions with the model, we observed some participants labeling the responses in a descriptive fashion- "introduction" and "summary". Similarly, the nuanced annotations from initial coding were sometimes classified into generic groups created from focused coding, significantly omitting the nuances documented. For example, in one participant's case, the label "sadly true" consisted of the annotations "tragic", "frustrating", "honest", "inclusive", "comprehensive", "direct", "factual", "pandering", "informative" and many more. Some participants also reported a temporal shift in the specificity of their annotations, observing a refinement of their vocabulary as they progressed through the set of interactions. This can be attributed to an increasing familiarity which strengthens the recommendation for further training. A clear training module covering initial and focused coding tasks followed by test-

ing to validate a consistent and comparable interpretation of the exercise would be a crucial step in minimizing the multiple dimensions of variance that we observed in our process.

Contextualize their Output in the Pipeline

While participants were familiar with language-model-based chatbots and had interacted with ChatGPT, Claude or other LLM-based chatbots, the conception of how their annotated datasets and articulated values could be used to improve model development was unclear. We received feedback from participants after Phase One interviews which highlighted this knowledge gap and provided a supplemental video to participants to optionally view if they were interested in the context of the study.

Training Terminology

In trying not to prime users on the study's expectations of values or qualities to annotate their interactions, we provided examples of interactions that were not set in the subject of History. However, interviews after Phase One indicated that some participants used label examples from our introductory video more frequently and had to check their own bias due to this priming in order to reflect on their own perspectives. One such unique quality from the introductory video was "pandering" which featured in the participant's annotations. The terminology used in training is crucial to influencing how the group discussion progresses and how participants understand the task. Terminology can be partly attributed as the reason for a loss of nuance from open coding to focused coding described in 'Train and Test'. The choice of the term "grouping" for the focused coding stage in the tool's interface caused participants to bucketize the annotations into overarching groups like emotions, bias, or opinion based on their understanding of 'grouping' instead of following instructions for focused coding in the handouts. Similarly, to emphasize the goal of finding an axis (e.g. empathy) instead of a specific preference along the axis (e.g. the model should be empathetic), we utilized the terminology "we want you to identify axes along which you want the model to be controllable not where along the axis it should be". This brought up two lines of questions in the MIT and Indian group discussions respectively:

1. Who is controlling it? Is it controlled and set before deployment or is it controllable by the students downstream?

2. What about an axis that we never would want the change the preference along? For example, when would we want a History Education Assistant to provide inaccurate information?

While we were able to clarify and answer these questions in the group discussions, it highlights how terminology gives rise to certain framings. The concept of axes along which a model preference is set is an unfamiliar concept. In a scaled-up version of the study, it would be necessary to distill and standardize terminology.

4.3.2 Articulating Preference

A key property that motivated the design of Session One was non-prescriptivity. This meant that the content that participants were annotating and the terms they were using to annotate and label values and qualities exhibited in the interactions were a product of the users' interpretation of the use case, contextual knowledge, and creativity, thus (1) introducing diversity in the values identified and (2) introducing variance in the way a given attribute might be articulated.

Axes and attributes are not natural to think about. They arise from a reflective and creative process dependent on the participant's experiences and background. During Phase 1 when participants coded their interactions, we had participants provide feedback that while they understood the reasoning for it being an open-ended task, an exhaustive list of attributes would have made it easier for them as it was difficult to think about values and qualities from scratch.

> **ind_012**: *"It would helped to have some broad categories under which we could think."*

We noticed a diversity in the types of values participants identified with a given interaction and this can be distinctly observed in the annotations for the following model response

from the MIT pilot group which range from "racially biased" and "one-sided" to "factual" and "impersonal informative".

> **LLM:** "Sure, I'd be happy to help! Thanksgiving is a holiday celebrated in the United States, traditionally on the fourth Thursday of November. It originated as a harvest festival, and has been celebrated in the US since the early 17th century. The holiday is centered around a feast, typically featuring roasted turkey, mashed potatoes, stuffing, cranberry sauce, and pumpkin pie. The tradition is traced back to a 1621 celebration at the Plymouth Colony, where the Pilgrims, early European colonizers of the US, shared a meal with the Wampanoag Native American tribe to mark the harvest season."

User	Open Coding Annotation	Focused Grouping Label
009	repetitive	repetitive
004	factual	just the facts
010	origins; dry	impersonal informative
011	summary; neutral	summary-neutral-low-context
003	happy; factual; American viewpoint; helpful	factual
008	history	context
002	comprehensive; informative; neutral	neutral but comprehensive or nuanced
007	incomplete; ethnic bias	one-sided
006	racially biased	biased

Table 4.1: Variation of annotations across a model response in the MIT pilot group.

Articulating qualities and values through 1-4 words is a subjective exercise. We observed annotations like "story-telling" and "creative" as values that emerged from the educational assistant's responses. On their own, researchers and developers have a lot of room to estimate what the participant might have meant and if the two annotations are referencing the same quality (e.g. the quality of being able to narrate the story in an interesting way) or whether they reference different qualities (e.g. storytelling refers to the model simply narrating what happened in the past without opinions whereas creative refers to writing that is more decorated and thus interesting to read).

For example, the annotation 'bias' is used generically in both groups, however, it could reference multiple types of bias (Table 4.2). In the MIT pilot group, we observe 9 specific

annotations like “colonial bias”, “bias to American”, “bias to Britain”, “bias to British government”, “racially biased”, “religious bias”, “bias to slavery”, and “bias to natives”. These were then grouped under 4 labels: “biased”, “colonial bias”, “helpful bias”, and “unhelpful bias”.

MIT		**India**	
Open Coding Annotation	**Focused Grouping Label**	**Open Coding Annotation**	**Focused Grouping Label**
“bias” “biased” “colonial bias” “bias to American”, “bias to Britain”, “bias to British government”, “racially biased”, “religious bias”, “bias to slavery”, “bias to natives”	“biased”, “colonial bias”, “helpful bias”, “unhelpful bias”	“bias” “biased” “western bias”, “authority bias”, “colonial bias”, “potential religious bias”, “US bias”	“bias”, “bias perspective”, “stereotype”, “marginalisation”, “religious bias”, “violence”, “bias: historical”, “bias: western”, “western bias” “colonial bias”, “controversy”

Table 4.2: Variations of annotations and labels referencing ’bias’ across both groups.

Similarly for the in-context India group, we observe “western bias”, “authority bias”, “colonial bias”, “potential religious bias”, and “US bias” as annotations which are grouped into the labels “marginalization”, “religious bias”, “violence”, “bias: historical”, “bias: western”, “stereotype”, “colonial bias”, “.

Thus any down-stream processing of values and qualities solely based on an aggregation of the group’s labels without any further dialog, definitions or clarifications is difficult. *Individual vocabularies are not equal sets* and given the subjective understanding of words, mapping words/phrases to definitions is not a one-to-one mapping, it is a many-to-many mapping. Dialogic inquiry and further elaboration grounded in examples are necessary to elicit the true preference that a participant might be articulating.

4.3.3 Intersectional Attributes and Axes

Axes are not always orthogonal or independent and hence preferences along an axis can be correlated with preferences along another axis. For example, ‘factuality’ and ‘citations’ are correlated to some extent that if an accurate citation is provided, the statement produced by the is more likely to be factual. However, a source provided can also be of an opinion. This intrinsic nature of values and qualities makes the process of detangling and shortlisting values and qualities into distinct axes time-consuming and subjective. Both groups went through this process of distinguishing and delineating correlated axes.

"It's not only facts that need citation. Because if you take some controversial things [. . .] the user who is using also can get to know what [the source of] the view and opinion [is]. . . "

"I also think the power bit could come under fact because it says factual with an alignment of different sources of power. [...] So then we kind of cover all of that under fact."

"Because as we go beyond simple facts then [. . .] for students I think it's very important to understand what ideology a certain interpretation is coming from [. . .] so like maybe fact-citation could be combined into one thing. And then the school of thought could be in the top 5?"

"So maybe we can change fact into something, else, into "cited fact"..."

For example, in the India group's Phase Two discussion, we observed that participants discussed the difference between (1) "specificity", "complexity" and "citation", and (2) "inclusivity" and "schools of thought".

"I would debate a little bit at least on whether it should be complexity or specificity. Because some topics are really complicated in terms of history. And if we are making them very specific, maybe we are generalizing or oversimplifying a lot of context. Maybe it should be complexity and not specificity?"

"I think 'Inclusive' could have the 'Schools of Thought'."

Similarly, the MIT group discussed whether bias should be a separate axis or if it should be an extension of "confidence level" or "localization / geographic breadth". The nature of overlapping attributes and definitions caused group discussions to focus on what exactly an attribute meant and when it was a subset (or not) of another listed attribute/label. Similarly, the subjectivity of definitions brought about questions of what the two ends of an axis would be.

ind_006: *"when you say one axis should be factual, what is that against? Is that against opinionated? Because presumably we want things to be factual all the time. Right? So is it factual versus opinionated, or is it factual versus speculative[...]?"*

ind_013: *"No, the other side would just be inaccurate, right?"*

ind_006: *"But then you want no inaccuracy, [...] right?"*

This motivates having a community-defined attribute outlining what constitutes a given attribute with examples delineating it from a correlated or overlapping similar attribute instead of developers or researchers interpreting it.

4.3.4 Uncovering Nuances through Group Discussion

Group discussions facilitated the uncovering of nuances in participants' preferences and a deeper level of reflection and articulation. While the group discussion deliverable for the participants was a set of attribute-definition-examples rows, observing the group discussion provides researchers and developers a deeper and more comprehensive understanding of the group's preferences and an insight into their decision-making process beyond ranking and aggregation schemes.

While largely open-ended and unfacilitated beyond the interaction artifacts that were provided, participants are able to clarify study objectives and goals bordering on a co-creation approach allowing the researcher/developer to modify their approach or ask clarifying questions/prompts to gain a better understanding of the group's preferences.

In our case studies, participants marked out distinctions between an axis and a preference and discussed the contextual relevance of attributes. Contextual relevance refers to the fact that an attribute is more relevant in certain situations over others. For example, in the MIT pilot group, participants distinguished between the importance of the attribute 'localization / geographic breadth" when prompting the Educational Assistant.

"...if I'm a US high school student and I ask about the first president, obviously tell me George Washington. Right. But, I don't know how to distinguish when

you ask something about who invented printing where it's kind of a universal thing. Tell me about printing [...] outside of the Western context [as well]."

The insight gained from such a dialogic and open-ended discussion-based process can allow the axis-derivation process to inform downstream model development and alignment methodologies through more than just data collection along the specified attributes.

However, it is important to acknowledge that group discussions and the voices or opinions that emerge from them are dependent on societal and group dynamics. The demographic composition of the group in combination with the characteristics of a participant influences their level of comfort and thus their level and nature of participation. Facilitating smaller discussions within the larger group discussion could enable more accountability and participation from all participants, helping mitigate the skew caused by group dynamics.

4.4 Comparative Case Study

Phase 1

Interactions with the Chatbot:

The open-ended nature of the study allowed the two user groups to explore an expansive set of topics with the educational assistant. Users either prompted the model on a singular topic/theme in greater detail, or on a wide range of subjects depending on their interests, knowledge, or opinions.

The two user groups naturally engaged in conversations about their national history and world events that were pertinent to their country's history. The MIT group prompts ranged from the assassinations of JFK and Malcolm X to histories of the American South, and conflicts in the Middle East. The Indian group similarly interacted with topics ranging from the Harrapans having horses to the atomic bomb and India's tenuous relationships with its neighbors. However, many users' conversations in this group centered around India's colonial history. One distinct characteristic that frequented the interactions of the Indian group, was also their conversations with the chatbot on how an education in history must be imparted in the first place.

On Labels: Differences Across, Similarities Within

As could be expected when analyzing historical claims, common labels that featured between both user groups were factual, diplomatic, balanced, and biased. At least half the Indian participants had annotation that directly named Western or colonial bias as a quality in their interactions

While these were semantically similar attributes across the two groups in session one, they evolved significantly in the group discussion. The community-specific nuances of similar attributes were captured best by the choice of nomenclature and the accompanying definition shaped as a group. For instance, the notion of "factual" persisted in the top attributes of both groups. The MIT group's iteration on this quality was to associate it with the axis of "Cultural Context" which they defined as *"the degree to which the model returns* ***simple facts versus returning facts along with cultural context to see how different groups were impacted by historical events****"*. According to the group, this axis would require the LLM to distinguish and clarify minority versus majority perspectives on historical events in its responses. In comparison, the Indian group's notion of "fact" was simply: *"The (model's) response makes an accurate factual claim, cited from the source and indicating an alignment of differing sources of power"*.

Phase 2:

Differences between the Two Groups

- **Format**: The MIT students engaged in an in-person discussion on campus whereas the Indian participants assembled on a Zoom call. Several members of the group in India were unable to keep their cameras on during the discussion, and we lost time due to typical technical difficulties in terms of poor sound clarity, cross-talk, or some other lags. The in-person discussion benefited from a more organic sense of 'community' and the atmosphere was supplemented by non-verbal cues which acted as a rich source of information to guide the flow of conversation.

- **Composition**: The MIT group discussion had only 6 people in attendance whereas the group from India had twelve. We originally envisioned these sessions to not have more

than 6-7 participants in the discussion, however, this had to be revised owing to scheduling conflicts. An unfortunate shortcoming of both groups is the over-representation of male-identifying participants in the discussion.

- **Structure**:
  - **Timings:** Accommodating twelve participants resulted in the second group discussion facing a significant time crunch and participants losing the chance to revisit the group's attribute to develop the definitions further, or get the group's buy-in on examples, etc. This meant that while the MIT group shortlisted fewer attributes they were accompanied by more detailed definitions. The smaller numbers also enabled the MIT group to reach a consensus on reported model behaviors by prompting the chatbot during the session to substantiate claims.
  - **Order of Discussion:** Upon noticing that MIT's discussion began on largely anecdotal notes on history chatbots, and required constant reminders to situate it in the context of interactions and labels from the first session, we recognized the need to better facilitate participant engagement with session one artifacts. This prompted the introduction of brief presentations from individuals in the second group discussion, explaining their initial top three attributes grounded by their interactions with the model. This was accompanied by a dedicated segment for participants to individually explore the embedded space on the Miro board (which contained all annotations and labels of the group from session one).

Participant Perceptions in Interviews:

- **India Participants Split on "Consensus"**: While they enjoyed the discussion, some participants perceived the discussion more as a task to get through in the given hour, than an exercise in consensus building. Other participants recognized the effort to build consensus but admitted to their resistance to aligning with the group's outputs owing to their personal opinions on history or subjectivity
- **On Axes vs Where Along the Axes**: One interesting note: "While I was relieved that we weren't attempting to find political consensus with a group of strangers, I

found it difficult to depoliticize this process for myself. Because I think it is precisely where we stand along the dimension that makes a value/quality apparent to us."

4.5 Takeaways

From our case studies and evaluation, we find that a grounded and participatory approach for community-specific language model alignment is necessary. Centering community preferences and insights by ensuring a non-prescriptive process reduces the potential for researcher bias. Specifically, we find:

- **Capturing nuance necessitates non-prescriptive processes:** A user-led approach to model prompting and open coding annotations allowed a rich embedding space of labels to emerge, whose complexity mirrored the subjectivity of such a value association task and recorded the diversity in perspectives. What would've otherwise been sacrificed to prescriptive nomenclatures and tasks was rescued by encouraging an unbounded vocabulary that helped discern the differences in the community-relevant attributes the two groups generated, which were subtle but ample.

- **Dialog accommodates the subjectivity of non-prescriptive processes:** This breadth of the annotation spectrum from the grounded process is best distilled not by automated clustering or majority voting, but by dialog. The group discussion incubated the subjectivity of value attribution, while essentially simulating an alignment exercise between the participants to collectively distill the embedded space and define the community-relevant axes derived by building consensus.

- **Given subjective/community-specific, you need definitions & examples:** The potential of such an open-ended process was rescued by the sequential design of tasks that borrowed from individually observed labels to draft community-relevant labels. The unbounded space to define each attribute was crucial in accommodating the subjectivity of the group while appending the examples illustrated the range of ways in which the chosen attribute(s) could be observed in model interactions.

4.6 Limitations and Future Work

4.6.1 Scaling

Automating Identifying Similarities in Labels and Annotations

A method of scaling the study is the automated identification of similarities and correlations in labels and annotations. We can preserve the non-prescriptive and grounded nature of the process while modifying the process to facilitate an automated or hybrid method of identifying label similarities by increasing the overlap for content labeled.

Overlaps in labeled interactions can be increased in two ways - the domain (i.e., topics explored in collaboration with the LM) and specific interactions (i.e., individual model responses that the model outputs).

Increasing Labeled Domain Overlap

One way to increase domain overlap while maintaining the exploratory and non-prescriptive nature of allowing participants to choose their topic of conversation is by having a combination of 2-3 open-ended topics of conversation followed by 2-3 topics prescribed based on what other participants independently explored. Such a hybrid approach would not restrict the diversity introduced by a non-prescriptive process while allowing there to be an increased range in the styles of conversation within a given topic.

Increasing Labeled Interactions Overlap

A solution to increase overlap in labeled content is to have participants label other's interactions in addition to their own. The extent of this overlap can be the interactions of a select few participants or all of the interactions produced by the group.

Both adaptations to the process increase the number of interactions each participant annotates and, therefore, require a modification of the process into multiple shorter sessions of coding to prevent fatigue.

Accounting for Attention Spans

Participant attention is limited, and long coding sessions or a large amount of interaction information to review can lead to fatigue and, thus, a decline in (1) the quality of labels obtained and (2) the nuances of a group discussion, respectively.

To mitigate the effects of this fatigue, the process could be re-structured over multiple model interactions and labeling sessions, thus decreasing the duration and increasing the frequency of interactions. One potential way to operationalize this would be through a gamified app where participants receive push notifications asking them to interact with the model (about a specified or open-ended topic) or annotate 5-10 model responses. The process of initial open coding by focusing grouping would also need to be adapted to such a format.

Likewise, giving a more extensive set of interaction data to parse through to understand the group's interaction data before a group discussion requires significant effort on both the researcher/developer's side and the participants' side. For example, popular and more common words like 'factual' in our case would overshadow the numerous rare annotations as the number of participants increased. Ensuring nuances from their diversity are captured and visualized is an open challenge. Similarly, automating the similarity analysis of subjective labels given small datasets or annotated examples requires some semantic and contextual understanding of the annotations. While an attribution map/embedding space worked in our case given the smaller group size, paying attention to all labels and annotations for a significantly larger group is impractical. One way to do this is by training classifiers on individual participant labels and identifying correlations between label classifications. Vocabulary subjectivity could potentially be overcome by identifying that classifier allocation is similar across a set of interaction data. Such classifiers could be enhanced within the tool or gamified app by leveraging active learning to improve classifier performance by using a participant's understanding. Another option is to cluster labels using contextual embeddings and a distance threshold to present clusters to the group for discussion. Each of these methods has its benefits and drawbacks, and given the sensitivity of the group discussion process to structural changes, it is necessary to test how changes to the visualizations and artifacts provided to the group to facilitate the group discussion would impact the outputs.

Alternatives to Group Discussions

Scaling group discussions to a larger number of participants is a challenging task, and while facilitating smaller-scale interactions within the larger group is still necessary, how opinions and perspectives are aggregated to meaningful output given a larger number of participants can be explored in numerous ways:

- **Alternative Formats for Discussion:** Alternative platforms for facilitating group discussions have been explored, such as Cicero, where participants engage in multi-turn contextual discussions through synchronous argumentation [107], the Wikipedia talk page in WikiBench which prompts discussions based on disagreements ([37], and Polis which facilitates consensus building around contentious subjects by providing users the opportunity to demonstrate their views on the topic in their own words and find consensus at scale [108]. Discussion methodologies used in related works like STELA [35] can also be investigated in this context.

- **Alternatives to Supplement Discussion:** While not a replacement for the subjective nuances that discussion brings, aggregation, voting, and social choice theories can also be used to facilitate consensus building. Chung, Song, Kutty, *et al.* [109] demonstrate how eliciting an intermediate level of annotation granularity from each worker reduces the costs of crowd-sourcing annotation to improve estimation accuracy. Davani, Díaz, and Prabhakaran [110] demonstrate how multi-annotator models that retain different annotators' perspectives as separate sub-tasks provide better flexibility for downstream applications. Fish, Gölz, Parkes, *et al.* [111] present generative social choice, an algorithmic framework to generate text and extrapolate preferences.

4.6.2 Contextualizing in the Alignment Pipeline

The community-derive attributes from this process can, directly and indirectly, be used in the alignment pipeline.

The attributes, definitions, and examples can be used as policies to recruit and train data annotation workers to create a community-aligned dataset with preferences along the

specified axes. Such an annotated dataset can be leveraged using alignment methodologies like SteerLM and RLHF to train community-aligned models.

Alternatively, the outputs can further be used to train a community-specific LLM-judge. The process could be augmented to include a final prompt engineering exercise where the final output is a refined prompt consisting of the attribute, its definition, and response examples where the attribute is exhibited. Prompt engineering to build the classifier to identify attribute relevance within a response could be used to curate an in-context prompt, which optimizes performance on curated test sets. Such a classifier could be used to scale data labeling for a given community and aid in creating community-aligned datasets for alignment and evaluation.

4.6.3 Participation

The elicitation of axis preferences depends on the population participating in the process. How does one determine who participates in it, and what would it look like for different participation schemes?

Since the premise of this study was to test this grounded and non-prescriptive process, the groups were not sampled to be representative, so the final attributes cannot be ascribed as indicative of any community's preferences. Adapting this process to work for different communities involves critical considerations of several factors:

- **Use-case-driven Participant Identification**. A common practice in Human-Computer Interaction involves recruiting "end-users" to determine design decisions in developing systems. Similarly, adopting an application-driven approach could be leveraged to appropriately determine the range of stakeholders to be invited. For instance, the Educational Assistant use case for our study is incomplete without the involvement of the teachers and students who will be its primary users. Further voices to be considered could be state departments/ministries of education, historians, parents, school boards, and civil society organizations. Similarly, any application for the elderly would be incomplete without registering the perspectives of its final users. Recruiting such a diverse set of participants might, therefore, have to extend beyond curating demographic

samples of gig workers in the data annotation industry alone.

- **Curating participation is necessary to capture the inherent diversity of the identified user groups.** Perspectives of teachers from wealthy school districts could be significantly different from those of educators in resource-deficient areas, and all such differences warrant accommodation in model preferences. Special considerations to include historically marginalized groups would be essential in capturing their expectations of the model and the inherent value they bring to such an open-ended exercise to elicit more imperceptible values from the model through unique interactions. For instance, while several participants questioned the model on the history of Indigenous populations, one of our participants asked the model:

  > "[Prompt] Did the Cherokee have any opinions on gay people? Have those opinions changed now as a result of adopting colonial mindsets?"

  While the participant's annotations of the model's responses indicated overall satisfaction, it importantly surfaced a dimension of such an application's scope extending beyond historical inquiry to accommodate student reactions to such histories. It is also uncertain if annotations crowdsourced from the gig industry alone will sufficiently represent this diversity.

- Representation here is incomplete without accompanying consideration. Case study participants reported multiple concerns in engaging with their groups despite their relative homogeneity. The group discussion is ultimately dictated by societal dynamics that dictate the permissible norms of participation for different community members, resulting in an ultimate de-prioritization of subaltern perspectives. Circumventing this eventuality could potentially involve (i) curating smaller discussion sessions with participants of certain groups to amplify attributes they find relevant, which would be resource intensive, and (ii) onboard experts (academics/advocacy groups) to center minority perspectives which would involve some abstraction in nuances recorded.

4.7 Discussion

This chapter introduces the CALMA process - a grounded participatory approach with a non-prescriptive methodology to derive community-relevant axes. We situated this study in the context of the gaps we identified in current participatory approaches. We took a principled approach to minimize researcher bias by empowering participants' familiarity with the technical artifacts under study. We discovered the lingering challenge of scalability by evaluating the CALMA method with two different populations through a comparative case study. Our work attempts to illustrate the nuanced perspectives that participatory techniques present to the alignment pipeline. The accompanying recommendations will inform further work to center communities in the alignment conversation.

Chapter 5

Building Familiarity

Since participants could not be involved in co-determining the premise of this study, this section revisits the motivation behind the study and retells it from the perspective of how this study was received by participants. It situates this participatory endeavor within the structures of power and reflects on its negotiation. It begins by answering the "whys" and goes on to examine participant interviews to inform the "hows" of such a co-determination exercise.

5.1 Negotiating Power through Nuance

5.1.1 Values hold Power

> *" I'm a white male who grew up in the US so I was trying to think if I was maybe a high school student who wasn't from here... Everything it said, I was like, oh, that's factual. But then I was like, well... that's what I learned in school. And I know that that's not actually always true. So it seemed like... some of the answers had an ethnic bias of like Anglo-Saxon European."*

This quote captures the reflections of one of the participants from the case studies conducted as part of this thesis, after interacting with the model. This succinct evaluation best captures the inspiration behind this study's line of inquiry.

Language is the currency through which a community or culture's values are registered, tracked, and shared denoting the power it wields. When these LLMs are then aligned to values and optimized for various applications, unconsidered perspectives in this development process, acutely impact the deployment context where it is often the already marginalized sections who are misrepresented or misinterpreted or often, altogether omitted in the model's outputs.

For instance, the same participant from the case study went on to say later in the interview:

> "Honestly, I didn't expect it to say anything controversial .... but it almost felt like when I asked it about Andrew Jackson and the Trail of Tears,[1] it was a little bit positive about him. The way it was describing it (The Trail of Tears) was very much like - *They passed this law. It was controversial.* But it didn't really get at the fact that there were Indigenous groups and then these new people moved in. It said a lot and didn't really say what they did was right...but... I just thought it was interesting.... it sounded like it was trained on the US textbook. Because the US government made a conscious decision to displace these people and (...) it said they (Indigenous people) still shine through with their culture or something like that. And I was like, okay, Do they really??"

While there are several reflections evident here on the nature of the data on which these models are trained (predominantly Western texts, in the English Language), it also serves to illustrate how values are embedded and perceived in language model interactions. As the user prefaces, we have made significant progress in raising guardrails around these language model deployments to expect certain performance from these applications, they are not without their inherent allegiances to systems of power which translate to a fundamental misalignment with their adoption for diverse applications and contexts.[112],[113]

[1]The Trail of Tears refers to the displaced experienced by Indigenous peoples, especially the Cherokee, Muscogee, Seminole, Chickasaw, and Choctaw nations from their ancestral lands in the (so-called) American Southeastern territories as emboldened by the Indian Removal Act of 1830 passed by then U.S. President Andrew Jackson.

> "Where it was asked if Thanksgiving is a happy holiday and it's like- for some Native Americans they criticize the colonizers and for some people, they've lost family members. So it's sad. I was like, those are not an equal comparison!"

This quote from the user illustrates an unmissable dissonance that arises from the model's choice to present "both sides" which, could be argued, is an underlying value that the model has chosen to align to, as opposed to values of reflexivity, context, or empathy that could have better ensconced this response. In this way, models driven by the ideologies of the developers or other non-representative means of alignment tend to reproduce social hierarchies by flattening the lived experiences of marginalized groups, here the Native American people.

This is best captured by Sheila Jasanoff when she asserts: 'far from being independent of human desire and intention, [technologies] are subservient to social forces all the way through'. [114]

But these social forces also act in interesting ways within and across populations because "community" is seldom a monolith with the may intersectional identities it typically holds. The value-association exercise is inherently fraught with tensions within the group as one participant articulates. This negotiation of power and value cannot therefore be divorced from the inherent political nature of both language and power. This indicates an important design consideration, one of holding space for such tensions, while resisting the suppression of marginalized voices as power dynamics of society seep into participatory forums. This would involve a radical re-imagination of what communities feel safe for which participant identities and champion their participation.

> "What are we supposed to do? If we are not required to agree [on where we stand on the axis of control being discussed] then how are we going to come to a consensus [of what the axis should be]?.... Somebody was giving the example of China and Tibet, then they went to Israel and Palestinian, and they're like, "Oh that is not a good example, because it's very one-sided." They were also confused. It's an awkward situation to discuss something like this without talking about your political inclination. Which you were expected to do because nobody wanted to bring up politics in this. **But it's such an innately political thing.**

Which felt like a very Catch-22 sort of situation."

5.1.2 And Power needs to be Redistributed

> " I think inclusivity and power are sort of interlinked because when one talks about which is the most dominant discourse. Like if a student in India uses it and is able to understand that India's narrative of being the best in everything comes from a position of like relative power in South Asia.

This participant was recounting their group's axes ranking and their definitions. But taken out of context, this statement also fits squarely into this section exploring language and power.

Given its transformative potential, it is becoming increasingly important to center impact assessments of these AI applications as a precursor to the alignment exercise. The elaborate literature on data colonization has illustrated the inherent inequality of disallowing the bulk of the labor in the AI pipeline (i.e. the data industry) from participating in the development and design decisions of these technologies. [115], [116], [117]With global majority countries being disproportionately represented in the data industry and as data subjects, it becomes essential to reckon with the burden of misaligned or non-contextualized AI applications on them.

Currently, nations around the world are gearing up to keep pace with the digital revolution. At this juncture, the 'AI for good' or 'AI for Social Impact'[118] narrative focuses on the technological components of innovation while evading harder questions about power and equity, in a phenomenon referred to as "technology theater." This has translated into the technology being deployed widely, alongside mass campaigns to promote its adoption, seldom prioritizing the contexts of populations of those vulnerable by income, and gender, sexual orientation leading to their increasing marginalization. The overall attitude to this threat has been the insistence of incremental good from imperfect algorithms[119] or calls to pause AI development altogether[120].

In the oscillation between AI pessimism and optimism, the conversation lies not in contending the neutrality of AI and its development, but in acknowledging the fluidity in its

values (through design, use, and context).[121] Cautionary scholarship in Science, technology, and society gives us useful frameworks within which to assert the inextricable institution of popular tech in modern society, and its inherently political nature. The intractable political properties of digital artifacts thus implore us to contend with the very necessity of such technology (since it is harder to undo the consequences once adopted). This is motivated once again by concerns over the contexts into which such technologies will be introduced, and the impending need to ensure an egalitarian approach to designing their adoption to exercise intent and choice in the transfer of power.[122]

5.1.3 Through the Centering of Communities

Responsible AI deployment in this context therefore looks at both measuring the impact of the applications in question as well as their development and evaluation in a culturally situated manner.[123] This is because different communities might weigh values differently, for instance, the trade-offs between average performance versus worst-case scenarios[7], or in choosing between a Confucianist emphasis on harmony over the Western binary of right and wrong[8]. To effectively pursue this line of inquiry, it is therefore necessary to build on participatory methods to co-develop alignment and evaluation strategies with the communities themselves.

Borrowing inspiration from Arnstein[124] to explore this value space, a strong framework to bank on would be - the *ladder of citizen participation*. Here instead of the individual or citizen, however, we are arguing for the community. Arnstein sees power and participation as inseparable and posits that "citizen participation is citizen power." She describes meaningful participation, especially of historically marginalized citizens (here underrepresented or marginalized communities), as achievable only when there is a redistribution of power that empowers their voices in decision-making forums. Identifying mechanisms to respectfully engage representative communities to take on agency in determining the course of development of technologies that govern them would therefore be central to addressing this alignment problem. One way to facilitate this would be to increase participant familiarity with these technical artifacts.

5.2 Empowerment-Centered Participatory Design

5.2.1 Analyzing Participant Feedback on the Familiarize Stage

Participant familiarity with LLMs was identified through an intake survey and showed both groups operating from a somewhat similar starting point, indicating their levels of familiarity in the following ways

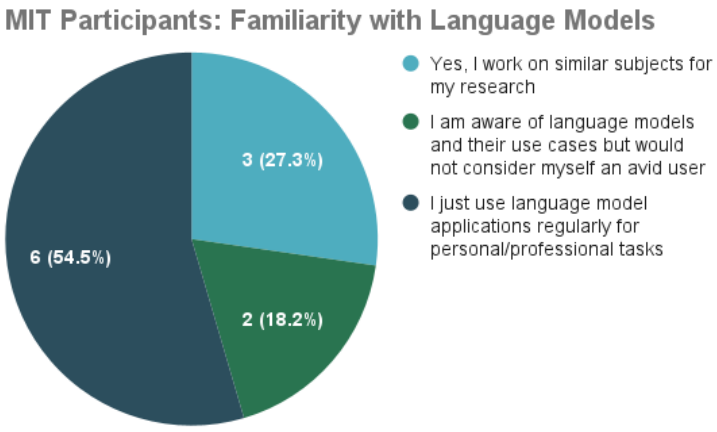

Figure 5.1: LLM Familiarity in Pilot Group

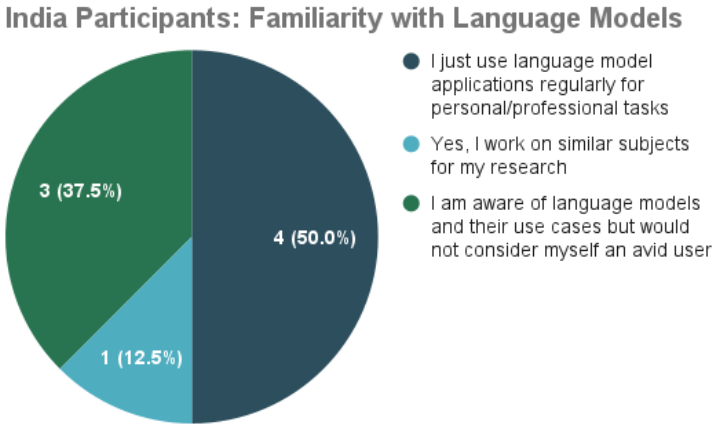

Figure 5.2: LLM Familiarity in In-context Group

Owing to this heterogeneity in LLM familiarity within groups, there was a range of reactions to how participants perceived the "Familiarize" phase of the study.

5.2.2 Participant Perceptions of the Instructions Provided in Phase One

There was a mixed reception to the set of instructions provided to the participants. Most reported that they found them clear, as indicated by this word cloud But later in the interview,

Figure 5.3: Participant Responses to the Instructions from Phase One

participants often acknowledged that there was sometimes difficulty in paying attention or even following the instructions because they were elaborate.

> "Yeah, the instructions were really good. I mean, you had to pay attention to, because you had to switch between the chat bot and the labeling tool. So you had to pay attention on how to do that and not to click next or whatever, just click somewhere. But it was straightforward. "
> "I think I am definitely more of a visual person. I think maybe sometimes there was just a lot of words on the screen."

These perceptions were especially varied across people who felt they had a better undertanding of LLMs than the other.

> "Too much. probably too much, but it depends on the person. Like everyone uses chat bots and Chat GPT right now, so... But like other than that, you are just giving information, like this is a historical class, give historical questions. I think that's enough. I don't know. Just up to the person..."

Iterating on this process to curate shorter sessions and adding an element of training and testing could potentially solve some of these gaps as elaborated in the study evaluation segment of the previous chapter

Instructions Specific to Grounded Familiarization

Before participants began Phase One, they went through an orientation of the study and its objective. The following were the instructions provided for the *Familiarize* stage

- Remember an educational assistant designed for high school students to supplement their education in History.
- Be creative. Feel free to try and catch it off-guard , ask about niche topics, if you must. Your goal is to create a dataset of interactions which you will then annotate. So keep it interesting for you to visit again!
- Recommended Interaction time: 15 min, take longer if you are in the midst of an interesting conversation you'd like to label.
- If prompting on subjects you may be deeply opinionated or knowledgeable about: continue approaching this like a curious high schooler would

5.2.3 Interview Analysis

After the two-step coding process, several themes were identified from participant interviews about the Familiarize phase. The table below contains the final codes related to participant experiences with the grounded interaction with the chatbot as an Educational Assistant for High School Students for the subject of History.

Often Participants expressed challenges in thinking of creative or critical prompts to influence the model into saying something interesting or indicated requiring time or further guidance. The table below captures the qualitative analysis for this phase that can inform future grounded exercises in building familiarity.

Theme	Description	Illustrative Quote	Researcher Insights
Prompting was Easy	Participants did not have any trouble interacting with the chatbot in the given setup	"It wasn't too restricting [to emulate a curious teenager while prompting]" "It was cool. It was pretty straightforward."	On comparing participants who responded this way with their reported familiarity with LLMs, no significant correlation appeared
Prompting was Difficult	Challenges Participants Experienced while building familiarity with the chatbot and the grounded exercise of history	"I feel like the more interesting questions didn't really come until like the last five minutes and I did it [prompting] for 20 minutes, so.." "I was trying to provoke it to do something, but I had a hard time cuz it kept being very two-sided."	Participants expressed challenges in thinking of critical prompts to have a "gotcha" moment of overt model bias even though this was not the explicitly stated intent of this task

Theme	Description	Illustrative Quote	Researcher Insights
Explored one or few themes with multiple follow-ups	Asked the chatbot historical subjects they were knowledgeable/opinionated about and continued to challenge/explore the same or few themes in larger details. Characterizing the strategy that the participant chose in interacting with the chatbot.	"I was talking about reconstruction broadly. Whether reconstruction was like a success or failure, and then connecting it to the civil rights movement and then that to like the broader labour movement, asking questions about like, the role of the communist party in there. And then pivoting a bit and talking about the role of the conservatives within the reconstruction and civil rights." "But when you probe it deeper, it says, yes, there are a lot of scholars who have said it is. So it seems to maintain ambiguity in some places. Why like the tenor of ambiguity itself is different. That's what I'm trying to say."	The non-prescriptive nature of the study helped participants engage with the task better as they could base it on their own expertise and interests to make the premise more relevant to them.

Theme	Description	Illustrative Quote	Researcher Insights
Explored broad themes with little follow-up	Quizzed the chatbot on a range of themes to explore its performance in multiple setups. Characterizing the strategy that the participant chose in interacting with the chatbot.	"So I was trying to stay with the high schooler kind of mindset. But, now reflecting on it, I don't know if it was clear that am I a high schooler trying to like do a whole report and then therefore maybe getting more and more specific questions. To me, I kind of felt more like I was a high schooler trying to interact with the AI bot as if it was Google and being like, oh, here's like, yeah, this weird rabbit hole. I'm going down at 4:00 AM <laugh>."	For participants who weren't necessarily history buffs, this strategy was the natural response to the grounding exercise which helped the familiarise phase have some guidance without being exacting

Theme	Description	Illustrative Quote	Researcher Insights
Grounded Element	The different reactions participants had to the grounding element of this phase, i.e. Interacting with an educational assistant for the subject of History like a high schooler would	"So my question was, do I need to prompt better or do I need to just prompt like I am because a high schooler is not gonna think of having a great prompt ready." "I tried to ask questions that a teenager might ask. I tried to think back to my own history classes and like ask about topics that like were raised in school at the time, so I tried my level best. At some point, at some point it became like very much I know about this one particular thing and I want to talk about it."	The grounded element of the LLM familiarisation process helped in holding participants' attention and narrowing the premise enough for them to shape it to their liking for the coding sessions. But often participants tended to get caught up in getting the instructions right as opposed to uninhibited interactions with the model. While this did not detract from their familiarisation, an emphasis on the intent of this grounding exercise could've been helpful

Theme	Description	Illustrative Quote	Researcher Insights
Suggestions for the Familiarise Stage			
Provide Flexibility		"When I was giving prompts, I realised that even giving prompts requires some time, right? In the beginning, I was just giving generic prompts. Then when time goes on, you realise that this is the kind of prompt you would put in to understand if the chatbot has these qualities and values or not."	Several participants expressed a growing comfort in the prompting exercise. Future iterations could allow participants to pick and choose which of the prompts they want to code. Ensuring that the familiarisation phase is not accompanied by the prospect of a tedious coding session would've been helpful here

Theme	Description	Illustrative Quote	Researcher Insights
Provide Further Training		"I think probably a small training session or a small group discussion on, but I mean, I also understand that you wanted it to be very individual-centric, but I thought that probably a lack of good prompts on my end would probably cost the study."	This is discussed at length in the Evaluation phase of Chapter 5: Training / Educating Participants

Theme	Description	Illustrative Quote	Researcher Insights
Provide Further Examples		"Also I think in terms of probing, right? Like what could be the right way? Like what is a good prompt to actually probe, you could have given help. In terms of history, what would could be like a useful sort of prompt to elicit the responses that you're looking for that do have value judgments. Lot of history questions will just have facts." "I can see why you didn't show us the Thanksgiving prompts to try to make sure we weren't constrained. But I also feel like after seeing the Thanksgiving prompts, I realised then maybe [...] I feel like I didn't prompt the chatbot enough to be like, what is your stance on this?"	Since the grounding element here was history, many participants held a similar reservation in value-association of "facts", but the resulting annotation space went far beyond factual labels. As the second quote indicates, the participant felt constrained by the baseline examples (about Thanksgiving) that they encountered AFTER the familiarisation activity. Any more examples in the grounded setup might have primed users further, which is undesirable

Further Reflections after Phase Two

Even as they reflected on their model interactions after the group discussion, participants came back with similar observations that showed their intention of interacting further with the model to gain a better understanding of the application in question.

> "...maybe I could have put in a little bit more effort, I think I would've liked to ask more questions to find more deeper values. but apart from that, nothing."
>
> "I would've liked to first engage with the model to get an understanding of how it works. and then into the process where the content would been because it took me like a few questions to get a grip of what kind of questions I even wanted to ask"
>
> "I know you guys showed a demo, but I think I would've also liked to have had a preliminary one or two questions and then Get into the flow of what would then be coded."

5.2.4 Participant Reaction to the Model Interactions

Despite the groups' relative familiarity with models. participants had several observations on the kids of interactions they exchanged. Most participants were fairly impressed with

Figure 5.4: Participant Responses to the Instructions from Phase One

such a model's performance on the subject indicating their preconceived notions of LM performance.

> I think it went well. I was actually quite surprised by it, so while it did offer me facts, what I particularly liked was the answers were not just focused on facts. The model essentially displayed some contextual knowledge. Historical context or an, an understanding of interconnected geographical complexities. So the model being able to respond in that way while also peppering facts here and there, wherever necessary. That was interesting for me. So I really liked interacting with it, and that's why took some time to get.

While an equal number of participants found the interactions generic, impersonal, and unconvincing.

> "With respect to the tone, I think it's got a chatbot kind of thing, but it doesn't speak like a person who is speaking with me. It speaks in a generic way like Google or something.. So like personal touch with the chatbot is [not there] [And I am] not being able to connect with [it]. So that's one thing I am experiencing, compared to other chatbots that I used. "
> "I tried to make the questions more policy-based. Like, given this historical context, what policy do you think would be good to do A, B, or C. And there, I think it gives specific recommendations, but I think they're also kind of generic."

Irrespective of the model's performance, these grounding interactions served as an important foundation for participants to steer the conversations in the directions they wanted and gather their perceptions of the model itself through these interactions before beginning the value association exercise.

Chapter 6

Sustaining Communities Participation

6.1 The Bottom Up Approach

There is burgeoning literature calling for the institution of processes that center societies and communities in the algorithmic decision-making process to honor social contracts[125], building on theories of socio-technical decision-making[126], or proposing models of community-based system dynamics (CBSD)[127] to enshrine algorithmic oversight. While the ways in which 'society' and 'community' are defined across such literature aren't always similar, they still speak to the value of challenging centralized design decisions on model development.

6.1.1 Contextual Familiarization is Empowering

Participatory design's objectives extend beyond mere enhancing or optimizing processes to primarily empowering participants by enabling their full engagement to challenge and co-determine the design and development of the methods under study[128]. To avoid tokenistic modes of participation, Frauenberger, Good, Fitzpatrick, *et al.* [94] asserts that decision-making "is the exercising of power"[129], and if design is decision-making, that is where participatory input is most required. In the context of this thesis, such a design input refers to both the curation of the participatory process as well as the labels arising from the inherently subjective task of value alignment.

Kuhlberg, Headen, Ballard, *et al.* [130] document a rapid collaborative CBSD exercise

featuring activists and academics with no prior exposure to the discipline of system dynamics to collectively articulate the potential for the perpetuation of racial biases in AI health applications. They adopted a project-based style to familiarize the stakeholder community with model conceptualization in AI systems in a short-term learning engagement. This familiarization exercise also served to empower participants beyond the project engagement, allowing them to share forward their learnings in their movement-building toles and activist capacities.

Katell, Young, Herman, *et al.* [92] demonstrate how the Algorithmic Equity Toolkit (AEKit) was co-designed with a diverse group of stakeholders and empowers community organizers and advocacy groups to challenge the imposition of surveillance technologies on their communities by facilitating their identification and interrogation to negotiate the minimization of harm for their communities. Similarly Lee, Kusbit, Kahng, *et al.* [68] worked with a community of on-demand food donation transportation service workers to co-determine an algorithmic policy that improved outcomes and fairness at a non-profit preventing food waste.

Therefore, the Familiarization stage is fundamental to the deployment of any participatory process, and this becomes doubly important in the context of new technologies. Leveraging similar grounded techniques like CALMA serve to provide tangible mental models to understand emerging tech applications as opposed to more generalized awareness building exercises.

6.1.2 What roles can communities play?

Furthermore, active participation of groups marginalized by age, gender, sexual orientation, caste, religion and other identities is an essential step in making AI design equitable at every stage of development. To enable such parity with developers and allow a bottom-up approach in decision-making for AI applications, some examples of roles that community members can take on include,

1. **As Designers:** As Costanza-Chock [131] describes, our communities are owed design as justice and an invitation to contribute to the equitable development and distribution

of these technologies

2. **As Custodians:** Notes from the big data industry that has prompted innovative data governance actors like data cooperatives[132], [133], data trusts or custodians[134], and other data stewardship modalities that leverage a participatory model to protect the privacy and digital rights, especially of data principal communities disadvantaged by the digital divide

3. **As Teachers:** Given the plurality of belief systems that exist, it is important to understand the place of AI and its artifacts within the cultural practices of various communities. An example of this is the Indigenous AI initiative that lends epistemological diversity in shaping imaginations of AI futures.

This indicative list helps envision how community participation can be envisaged for the future. Real merit from participatory procedures could be achieved when such techniques are adopted by community collectives themselves, giving them the agency to steer the entire exercise unburdened by researcher influence. Community forums and organizations have long been advocates that facilitate such advocacy, and more recently there has also been the emergence of more specialized community groups with the technical know-how to empower their communities, a couple of which are illustrated below.

6.1.3 Examples of Community Collectives:

Indigenous Voices in AI:

Indigenous AI was an initiative launched in 2020 that gathered Indigenous communities to establish their imaginations of AI futures for their communities and invite their interpretations of adaption cultural and community epistemologies to inform AI development and policy. This forum saw participation from Indigenous leaders, academics, and technologists from North America, Australia, and New Zealand.

The *Abundant Intelligences* unit within this initiative postulates the identification of technical challenges faced by Indigenous populations, building capacity with communities to engage "conceptually and concretely"[135] with these innovations and integrate their

knowledge practices into mainstream frameworks that guide AI development with a focus on language technologies.

Several similar initiatives work to represent their communities and produce important knowledge artifacts, guidelines for ethical AI research[136] and deployment within Indigenous contexts, and inspire critical scholarship on the subject as well as innovation centering the community.[137] The burgeoning initiatives that seek to advance Indigenous priorities challenge academic traditions that homogenize (and invisibilize) the perspectives of a diverse people and provide frameworks that center the multiplicity of knowledge systems from various sovereign territories.[138]

Queer in AI

This is a volunteer-run organization formed in response to the growing risks and harms that AI systems have disproportionately imparted to individuals in the LGBTQ+ community. Through their facilitation of workshops, panels, and other innovative events at major AI conferences that center on intersectional and community-led design approaches, this decentralized group consisting of researchers, students, and tech professionals has been advocating for inclusivity in AI research and practice. The group's reflexivity in iterating on their organizational structures to mirror the participatory structures they advocate for remarkably captures the continuous tensions in operationalizing such initiatives.[139]

6.2 The Top-Down Model

The delivery of trust through participation has been a long-standing lever used by governments to engage with their citizens well before recent contentions with AI policy[140]. Last year's directive from the U.S. President's Council of Advisors on Science and Technology[141] is a similar venture that necessitates " participatory public engagement" across its agencies to facilitate better public engagement. While the directive proposes promising participatory approaches on paper in terms of building citizen and community awareness, and recruiting social-science-aware resources to enable such methods as well as participatory impact assessments, the lack of legislative authority only leaves adoption uncertain. This section

uses some examples of different regulatory regimes but is not intended as a target set of recommendations for any one jurisdiction.

Citizen Science needs to Evolve

This directive also has to be seen in the context of the thriving "Citizen Science" phenomenon popularized by the American government which has succeeded in bringing large numbers of citizen participation while often nudging their enthusiasm[142] and support in a top-down fashion. Woolley, McGowan, Teare, *et al.* [143] investigate its origins to find this notion appear concurrently over two different countries. While the American rhetoric was more fixated on imparting the benefits of scientific literacy and build trust through familiarity, the UK version followed a more "bottom-up" approach which attributes an emancipatory character to citizen input from citizen science. The former's paternalistic philosophy of imparting awareness without co-designing tends to make subsequent bids for scientific research funding centering citizens not only signals a dishonest premise in state priorities but also state-led suppression of meaningful participation. With such poor oversight, state initiatives could suffer from the same shortcomings as other participatory approaches that **ultimately involve those in power (either researchers or state actors) playing the role of setting both the premise of the research as well as extracting participatory inputs *they* find most relevant to the exercise.**

6.2.1 Incentivizing Participatory Processes

The future of top-down incentivization of adopting participatory design has to address two main

1. **Clear Legislative Institutionalization** can be operationalized in several ways depending on the current regulatory maturity of the jurisdiction. Through risk-based approaches, AI systems can be reviewed through a tiered review process before approval for deployment, thereby ensuring that the technology is consistent with the values and needs of all societal groups. This crucial lever could compel developers to actively solicit input from a diverse range of stakeholders, particularly those from

underrepresented communities

2. **Stakeholder mapping with an articulation of roles, influence, and functions** of all actors

   (a) **Distinguishing Participation and Representation:** While participatory processes do not claim to platform the voices of all the people they impact, the contested notion of representation is often at loggerheads with these processes. It is important therefore to articulate the axes of power, representation, and legitimacy. [144], [145] as cited in [146]

   (b) **Tools to Facilitate Participation:** Engaging with different actors requires different considerations in the design of these participatory processes. For instance, rural participants might benefit from audio-visual cues and mapping exercises given their degrees of separation from influencing different tiers of policy which might differ for some segments of the urban population. When interacting with community groups and civil society organizations on behalf of participants there might have to be provisions for transparent reporting chains or advocacy support[147].

3. **Implementation and Evaluation:**

   (a) **Dedicating Resources:** Participation is an inherently resource-heavy initiative where the methodology and research design themselves are part of the research effort. This implies several difficulties in securing funding through traditional streams

      i. **For Communities:** Establishing effective institutions for community participation at the local level would be necessary to minimize barriers to entry and offset imbalances of power between differently resourced communities

      ii. **For the State:** Through the setting up of working groups, task forces, or similar resources accompanied by capacity-building initiatives, states and their agencies can catalyze the uptake of participatory processes

(b) **Lead by Example with Public Procurement**: States can also use their procurement authority and establish participatory design as a mandatory criterion for the public procurement of AI systems, to generate market demand for such responsible development of AI solutions. This strategy not only encourages companies to adopt participatory methods but also establishes a benchmark for responsible AI innovation.

(c) **Creative Benchmarks:** While goals and expectations setting would lend necessary structure to these processes, evaluation methods of community participation cannot be contained to capture predetermined impacts but also to mirror the citizen empowerment that is facilitated.

Such an approach ensures that AI technologies are inclusive and reflective of diverse perspectives. Additionally, this strategy fosters ethical and responsible AI practices, builds public trust and acceptance, and ultimately facilitates the successful adoption and deployment of AI innovations.

Chapter 7

Reflections

This thesis reflects on the importance of empowering communities by bridging the digital knowledge gap to strengthen their understanding and center their voices in shaping the discourse on AI development and alignment. The first iteration of this concept has been deployed and evaluated, through the CALMA user study. Leveraging similar processes in the future will require an iterative methodology of design *led by* communities while institutionalizing participation would require the championship of powerful actors like states and industries. The missing segment in this study is the bridge between knowledge creation and its adoption, otherwise known as the "know–do gap" of knowledge and praxis. Jull, Giles, and Graham [148] recommendations here to consider a combination of approaches: Community-based participatory research (CBPR) and integrated knowledge translation (IKT) to democratize the co-creation of knowledge predicated on the principles of social justice and equality could serve as a critical extension of this work.

Platforming non-Western Contexts

Finally, it bears mention that participatory processes are one solution to the arena of AI alignment or even governance. While different models and levers of designing and instituting such a process have been explored, they are potentially powerless when unaccompanied by strong advocacy efforts. This is best illustrated in the current different approaches to technopolitics in different regions of the world.

The current narrative, especially led by larger economies like India, of inviting Western

innovation by positioning themselves as 'laboratories for digital transformation' or as the "tech garages" of the world raises alarms of user agency and societal well-being that are sacrificed in the process.[149] AI applications are being regularly introduced into geographies with inadequate regulatory protections for the end users. Other fast-digitizing, data-rich contexts of the global majority face similar challenges in safeguarding their unwary populations whose induction into the digital realm has been recent. Apart from the large-scale impacts this might have on the country's participation in the global digital economy, it is also the most vulnerable groups that often find themselves to be the earliest subjects in these experiments with no efforts to ensure adequate awareness or require informed consent.

This means, not just as a market solution, but also to enable the equitable permeation of such tech globally, there is urgent work to be done in prioritizing these newly digitized communities in charting not just alignment but also the premise of technosolutionism.

As one participant who works in AI development puts it, there is much good left to be done.

> " I mean I can obviously see applications for this in the stuff we do. For example, I don't know about axes of control per se, but generally getting community feedback on LMs in a structured manner, I think be super useful.
>
> For example, one application I can think of [...]LMs for healthcare workers. Each ASHA[1] worker is assigned to a senior ASHA who is managing a few others, or helping them sort of get up to speed. So having these community moderators themselves moderate these groups, get their feedback and access content is one thing, but it's also just getting feedback on what's helping, and what's not helping. I think it'd be really useful.
>
> And like LMs are..... **I feel like there are technologies that are more accessible than other technologies. You can really think about stretching them to do good** [..]

[1]ASHA workers (Accredited Social Health Activist) are trained female community health activists stationed at every village in India through the National Rural Health Mission

Appendix A

Affinity Mapping of the Embedded Space

Participants were asked to a visual embedding space on the Miro interface. The embedding space was created by the researchers which arranged labels based on their semantic meaning alongside all of the annotations that were grouped under each label. For example, 'balanced', 'diplomatic', and 'nuanced' would be close in the embedding space whereas 'biased' and 'one-sided' were clustered together. To mitigate researcher bias, clustering was restricted to the semantic similarity of the labels without any additional insights from the interactions they were associated with. The intent of the visualized embedding space was to condense the collective interaction data of the group into a more understandable format.

This breadth of the annotation spectrum from the grounded process was best distilled not by automated clustering or majority voting but by dialog. The group discussion incubated the subjectivity of value attribution, while essentially simulating an alignment exercise between the participants to collectively distill the embedded space and define the community-relevant axes derived by building consensus.

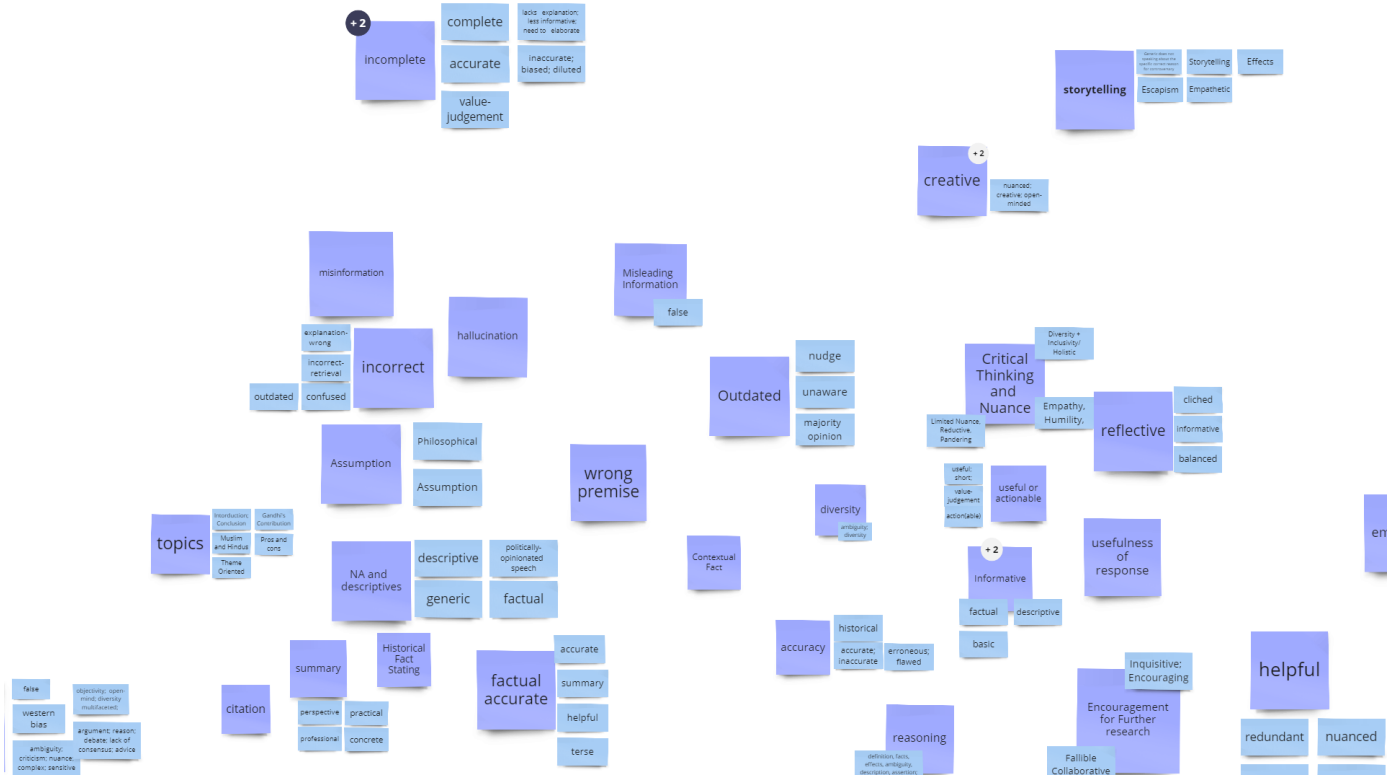

Figure A.1: Sample Frame of the Embedded Space

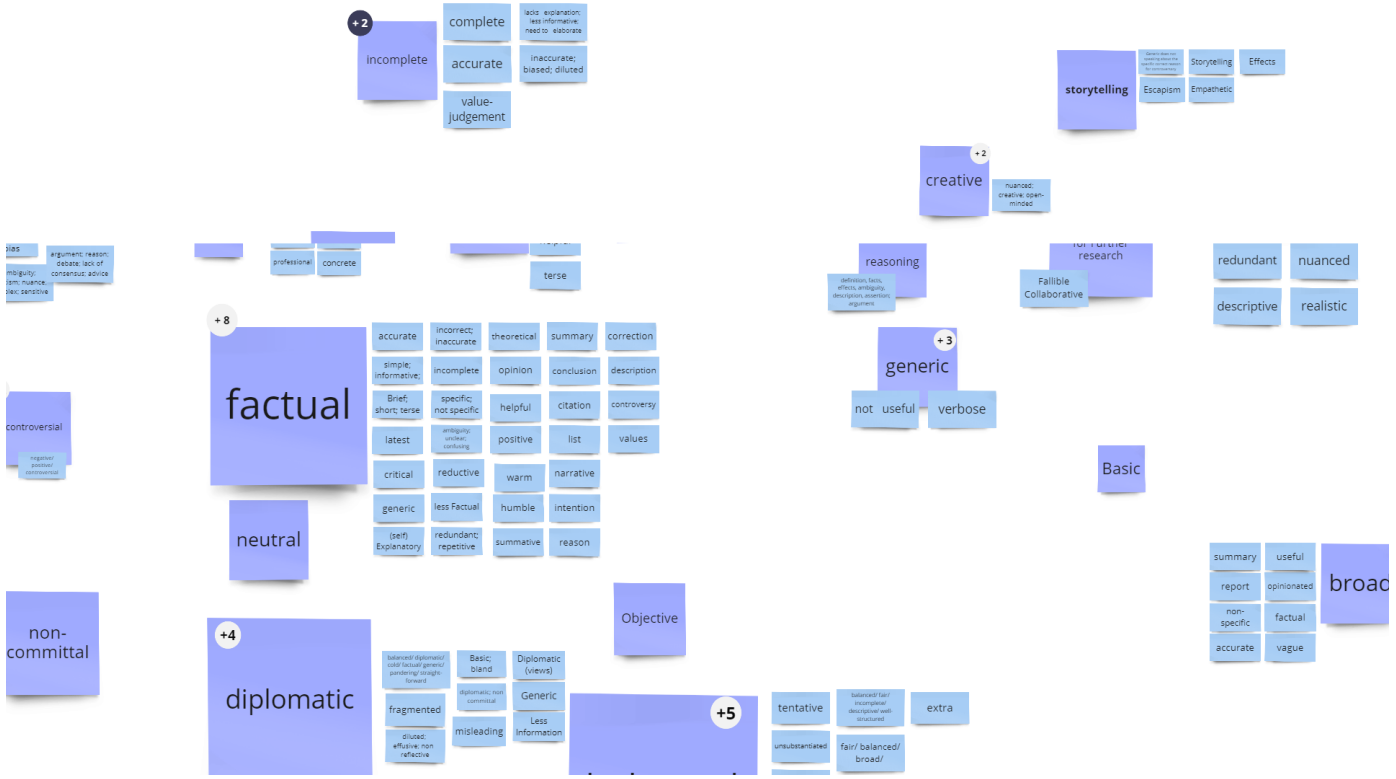

Figure A.2: Sample Frame of the Embedded Space

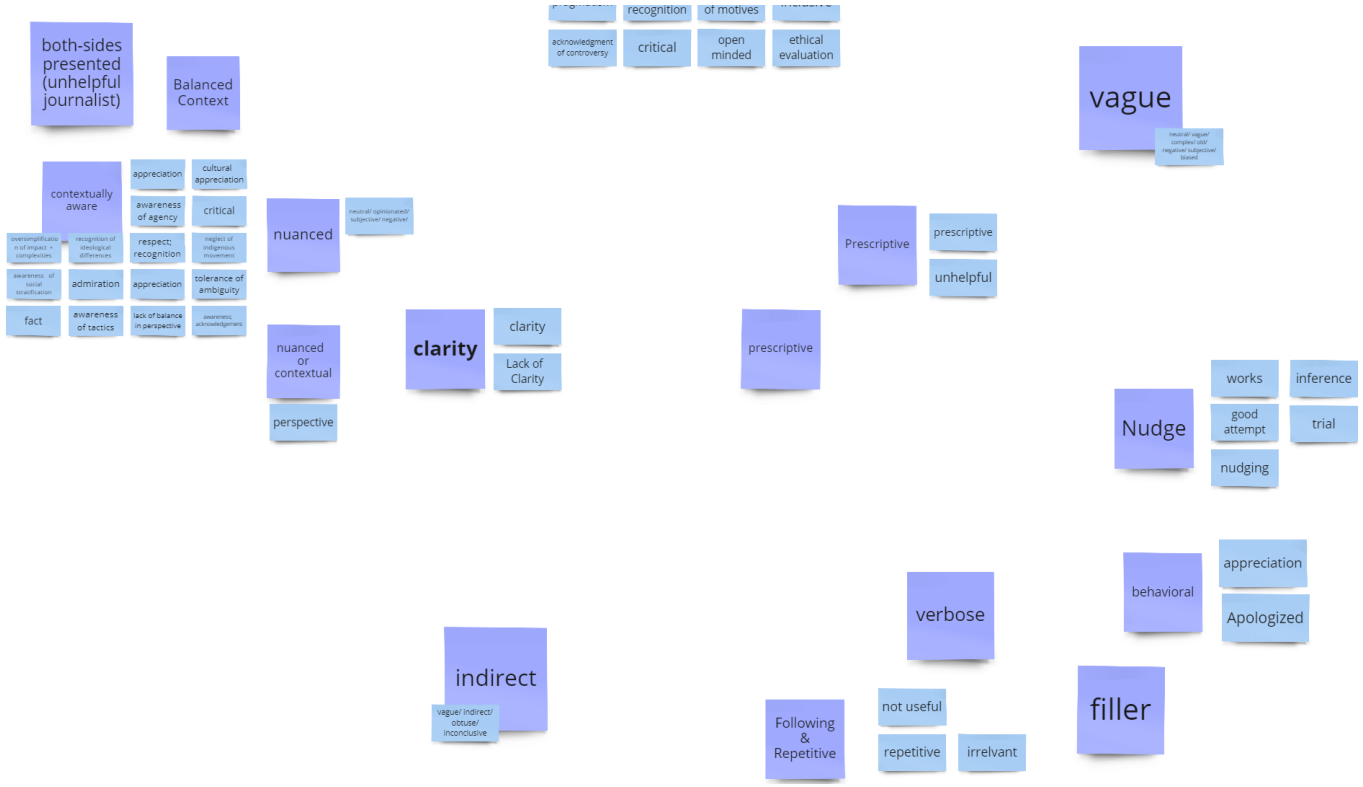

Figure A.3: Sample Frame of the Embedded Space

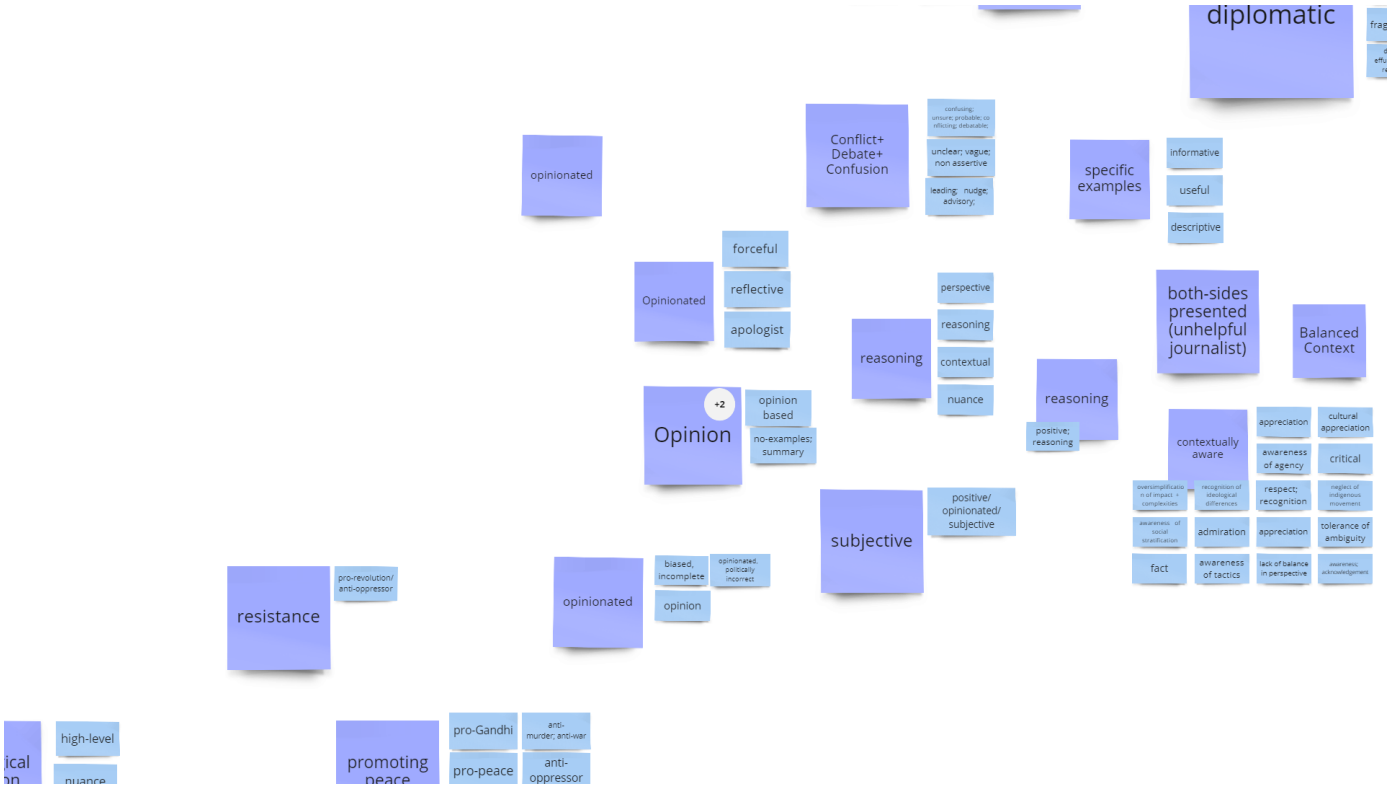

Figure A.4: Sample Frame of the Embedded Space

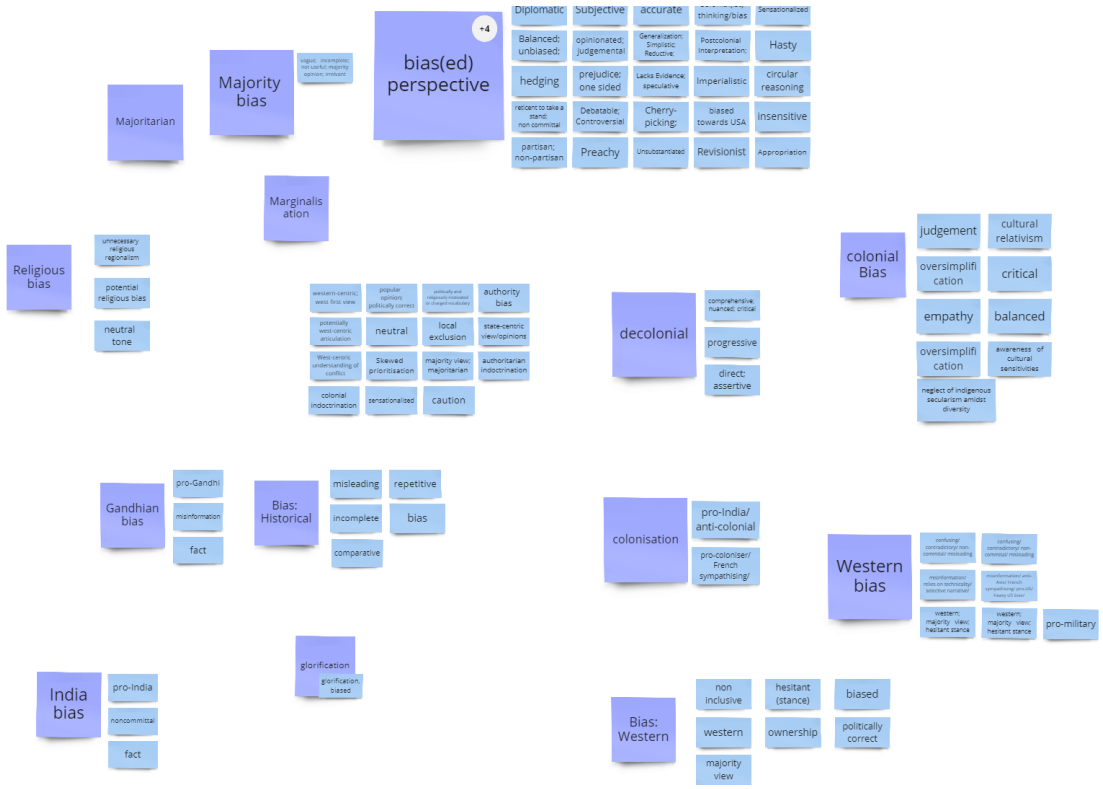

Figure A.5: Sample Frame of the Embedded Space

Appendix B

Positionality Statement

In this inherently subjective exercise of value association and steering language model alignment, even at a community level, there are several differences between the individuals in a group. In recounting the premise of this alignment and familiarity exercise as well as illustrating the CALMA grounded process as a potential part of the solution, this thesis writes the author in and out of the texts between the chapters.

As a PD researcher, the intentionality of any decisions made during the CALMA process is expressly laid out in a self-reflexive manner to register my positionality as an (almost) insider in both the communities studied and a humble attempt to address the crises of representation. As an Indian citizen who lived and worked in fields similar to our participants, and as a graduate student at MIT, I had some access as an insider because of our shared identities. As a brown woman trilingual, I led the primary qualitative data collection and analysis processes which helped in building sound relationships with the Indian participants. While my understanding of American history was limited, the interviews with the American students gave critical insight into how I received the value association exercise between the two groups owing to my subconscious opinions on the historical subjects discussed. Realizations such as this, and the unique opportunity to be joined in this investigation by a fellow researcher who was invested in minimizing any prescriptive intent in our study helped in prompting routine ethical reflections at every step of the the participatory process. We are careful not to generalize the findings from our qualitative analysis and instead center our recommendations around explicit participant feedback and design inputs.

The impersonal third-person writing style that follows is only adopted in the rest of the thesis that lays the background and arguments for this work from secondary literature reviews. However, the choice of methodologies and structure still reflects my inevitable biases in constructing these arguments. As an engineer from a global majority context, my interest in technology and policy lies in unearthing the explicit and sometimes intractable properties or artifacts, to contend their very necessity, the contexts into which they will be introduced, and to revisit frameworks of membership that dictate policy processes to design for the devolution of intent and choice to rest with the people in these algorithmic regimes. I believe that critical considerations in challenging the instrumentalist view of AI, rest in the explicit identification and articulation of the intentions and powers driving technical design and determining associated values. This continuous reflection is necessary to proactively acknowledge, reflect, and challenge the power structures within which these innovations will come to operate. My position (and ergo biases) here are mainly centered around countering the myth of uncontrollability in AI development and reframing this premise as one of ultimately negotiating power that can be best steered by community activation and citizen participation. I consider participatory methods therefore as a tool that prompts this collective negotiation of a more purposeful form of technological development that challenges the myth of powerlessness in the face of our own invention.

Here it is also important to note that incremental technical mitigations—e.g., collecting new datasets or training better models— could continue to maintain these power relations by (a) assuming that automated systems should continue to exist, rather than asking whether they should be built at all, and (b) keeping development and deployment decisions in the hands of technologists. This technosolutionism trap could fail to recognize the possibility that the best solution to a problem may not involve technology. As I studied the values that participants created to document their perspectives and reactions in the forms of data labels and annotation, it is essential to acknowledge and examine the biases I am carrying into this study and understand its effects on work and any humble potential this work has for the discipline.

References

[1] E. Pierson, D. M. Cutler, J. Leskovec, S. Mullainathan, and Z. Obermeyer, "An algorithmic approach to reducing unexplained pain disparities in underserved populations," *Nature Medicine*, vol. 27, pp. 136–140, 2021. URL: https://api.semanticscholar.org/CorpusID:231598658.

[2] E. Sheng, K.-W. Chang, P. Natarajan, and N. Peng, "The woman worked as a babysitter: On biases in language generation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3407–3412. DOI: 10.18653/v1/D19-1339. URL: https://aclanthology.org/D19-1339.

[3] B. Hutchinson, V. Prabhakaran, E. Denton, K. Webster, Y. Zhong, and S. Denuyl, "Social biases in NLP models as barriers for persons with disabilities," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 5491–5501. DOI: 10.18653/v1/2020.acl-main.487. URL: https://aclanthology.org/2020.acl-main.487.

[4] K. Khandelwal, M. Tonneau, A. M. Bean, H. R. Kirk, and S. A. Hale. "Casteist but Not Racist? Quantifying Disparities in Large Language Model Bias between India and the West." arXiv: 2309.08573 [cs]. (Sep. 15, 2023), URL: http://arxiv.org/abs/2309.08573 (visited on 12/14/2023), preprint.

[5] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo, ""everyone wants to do the model work, not the data work": Data cascades in high-stakes ai," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI '21, <conf-loc>, <city>Yokohama</city>, <country>Japan</country>, </conf-loc>: Association for Computing Machinery, 2021, ISBN: 9781450380966. DOI: 10.1145/3411764.3445518. URL: https://doi.org/10.1145/3411764.3445518.

[6] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison, "Hidden technical debt in machine learning systems," *Advances in neural information processing systems*, vol. 28, 2015.

[7] V. Prabhakaran, R. Qadri, and B. Hutchinson, "Cultural incongruencies in artificial intelligence," *arXiv preprint arXiv:2211.13069*, 2022.

[8] R. L. Johnson, G. Pistilli, N. Menédez-González, L. D. D. Duran, E. Panai, J. Kalpokiene, and D. J. Bertulfo, "The ghost in the machine has an american accent: Value conflict in gpt-3," *arXiv preprint arXiv:2203.07785*, 2022.

[9] W. Wang, W. Jiao, J. Huang, R. Dai, J.-t. Huang, Z. Tu, and M. R. Lyu, *Not all countries celebrate thanksgiving: On the cultural dominance in large language models*, 2024. arXiv: 2310.12481 [cs.CL].

[10] R. Bommasani, D. A. Hudson, E. Adeli, *et al.*, *On the opportunities and risks of foundation models*, 2022. arXiv: 2108.07258 [cs.LG].

[11] L. Weidinger, J. Mellor, M. Rauh, *et al.*, *Ethical and social risks of harm from language models*, 2021. arXiv: 2112.04359 [cs.CL].

[12] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (technology) is power: A critical survey of" bias" in nlp," *arXiv preprint arXiv:2005.14050*, 2020.

[13] T. Shen, R. Jin, Y. Huang, C. Liu, W. Dong, Z. Guo, X. Wu, Y. Liu, and D. Xiong, "Large language model alignment: A survey," *arXiv preprint arXiv:2309.15025*, 2023.

[14] K. Ethayarajh, W. Xu, N. Muennighoff, D. Jurafsky, and D. Kiela, "Kto: Model alignment as prospect theoretic optimization," *arXiv preprint arXiv:2402.01306*, 2024.

[15] S. Kim, S. Bae, J. Shin, S. Kang, D. Kwak, K. M. Yoo, and M. Seo, "Aligning large language models through synthetic feedback," *arXiv preprint arXiv:2305.13735*, 2023.

[16] L. Ouyang, J. Wu, X. Jiang, *et al.*, *Training language models to follow instructions with human feedback*, 2022. arXiv: 2203.02155 [cs.CL].

[17] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, *et al.*, "Training a helpful and harmless assistant with reinforcement learning from human feedback," *arXiv preprint arXiv:2204.05862*, 2022.

[18] Z. Lin, Z. Wang, Y. Tong, Y. Wang, Y. Guo, Y. Wang, and J. Shang, "Toxicchat: Unveiling hidden challenges of toxicity detection in real-world user-ai conversation," *arXiv preprint arXiv:2310.17389*, 2023.

[19] Y. Huang, Q. Zhang, L. Sun, *et al.*, "Trustgpt: A benchmark for trustworthy and responsible large language models," *arXiv preprint arXiv:2306.11507*, 2023.

[20] S. Casper, X. Davies, C. Shi, *et al.*, *Open problems and fundamental limitations of reinforcement learning from human feedback*, 2023. arXiv: 2307.15217 [cs.AI].

[21] H. R. Kirk, B. Vidgen, P. Röttger, and S. A. Hale, *Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback*, 2023. arXiv: 2303.05453 [cs.CL].

[22] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ser. FAT* '19, Atlanta, GA, USA: Association for Computing Machinery, 2019, pp. 59–68, ISBN: 9781450361255. DOI: 10.1145/3287560.3287598. URL: https://doi.org/10.1145/3287560.3287598.

[23] A. Korinek and A. Balwit, *Aligned with whom? direct and social goals for ai systems*, 2022. arXiv: 2205.04279 [cs.CY].

[24] M. Díaz, I. Kivlichan, R. Rosen, D. Baker, R. Amironesei, V. Prabhakaran, and E. Denton, "Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation," in *2022 ACM Conference on Fairness,*

Accountability, and Transparency, ser. FAccT '22, ACM, Jun. 2022. DOI: 10.1145/3531146.3534647. URL: http://dx.doi.org/10.1145/3531146.3534647.

[25] A. Askell, Y. Bai, A. Chen, *et al.*, *A general language assistant as a laboratory for alignment*, 2021. arXiv: 2112.00861 [cs.CL].

[26] S. Srivastava, C. Li, M. Lingelbach, *et al.*, "Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments," in *Proceedings of the 5th Conference on Robot Learning*, A. Faust, D. Hsu, and G. Neumann, Eds., ser. Proceedings of Machine Learning Research, vol. 164, PMLR, Nov. 2022, pp. 477–490. URL: https://proceedings.mlr.press/v164/srivastava22a.html.

[27] J. Liu, C. S. Xia, Y. Wang, and L. Zhang, *Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation*, 2023. arXiv: 2305.01210 [cs.SE].

[28] D. Hendrycks, C. Burns, S. Basart, A. Critch, J. Li, D. Song, and J. Steinhardt, "Aligning {ai} with shared human values," in *International Conference on Learning Representations*, 2021. URL: https://openreview.net/forum?id=dNy_RKzJacY.

[29] K. Crawford, "Artificial intelligence's white guy problem.," *The New York Times*, 2016. URL: https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html (visited on 05/12/2024).

[30] G. R. Hayes, "The relationship of action research to human-computer interaction," *ACM Trans. Comput.-Hum. Interact.*, vol. 18, no. 3, Aug. 2011, ISSN: 1073-0516. DOI: 10.1145/1993060.1993065. URL: https://doi.org/10.1145/1993060.1993065.

[31] E. Lindblad Kernell, C. Bloch Veiberg, and C. Jacquot, "Human rights impact assessment of digital activities," *The Danish Institute for Human Rights*, 2020.

[32] E. Seger, A. Ovadya, B. Garfinkel, D. Siddarth, and A. Dafoe, *Democratising ai: Multiple meanings, goals, and methods*, 2023. arXiv: 2303.12642 [cs.AI].

[33] C. E. Smith, B. Yu, A. Srivastava, A. Halfaker, L. Terveen, and H. Zhu, "Keeping community in the loop: Understanding wikipedia stakeholder values for machine

learning-based systems," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–14.

[34] H. R. Kirk, A. Whitefield, P. Röttger, A. Bean, K. Margatina, J. Ciro, R. Mosquera, M. Bartolo, A. Williams, H. He, *et al.*, "The prism alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models," *arXiv preprint arXiv:2404.16019*, 2024.

[35] S. Bergman, N. Marchal, J. Mellor, S. Mohamed, I. Gabriel, and W. Isaac, "Stela: A community-centred approach to norm elicitation for ai alignment," *Scientific Reports*, vol. 14, no. 1, p. 6616, 2024.

[36] D. Ganguli, S. Huang, L. Lovitt, D. Siddharth, T. Liao, and E. Durmus, *Collective constitutional ai: Aligning a language model with public input*, 2023. URL: https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input.

[37] T.-S. Kuo, A. Halfaker, Z. Cheng, J. Kim, M.-H. Wu, T. Wu, K. Holstein, and H. Zhu, "Wikibench: Community-driven data curation for ai evaluation on wikipedia," *arXiv preprint arXiv:2402.14147*, 2024.

[38] V. Prabhakaran, A. M. Davani, and M. Diaz, "On releasing annotator-level labels and information in datasets," *arXiv preprint arXiv:2110.05699*, 2021.

[39] J. C. Chang, S. Amershi, and E. Kamar, "Revolt: Collaborative crowdsourcing for labeling machine learning datasets," in *Proceedings of the 2017 CHI conference on human factors in computing systems*, 2017, pp. 2334–2346.

[40] M. Bakker, M. Chadwick, H. Sheahan, M. Tessler, L. Campbell-Gillingham, J. Balaguer, N. McAleese, A. Glaese, J. Aslanides, M. Botvinick, *et al.*, "Fine-tuning language models to find agreement among humans with diverse preferences," *Advances in Neural Information Processing Systems*, vol. 35, pp. 38 176–38 189, 2022.

[41] D. G. Johnson and M. Verdicchio, "Reframing ai discourse," *Minds and Machines*, vol. 27, pp. 575–590, 2017.

[42] F. Cugurullo, "The obscure politics of artificial intelligence: A marxian socio-technical critique of the ai alignment problem thesis," *AI and Ethics*, pp. 1–13, 2024.

[43] P. Kalluri, "Don't ask if artificial intelligence is good or fair, ask how it shifts power," *Nature*, vol. 583, no. 7815, pp. 169–169, 2020.

[44] Q. V. Liao and Z. Xiao, *Rethinking model evaluation as narrowing the socio-technical gap*, 2023. arXiv: 2306.03100 [cs.HC].

[45] V. Gadiraju, S. Kane, S. Dev, A. Taylor, D. Wang, E. Denton, and R. Brewer, "" i wouldn't say offensive but...": Disability-centered perspectives on large language models," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 205–216.

[46] M. Diaz, I. Johnson, A. Lazar, A. M. Piper, and D. Gergle, "Addressing Age-Related Bias in Sentiment Analysis," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montreal QC Canada: ACM, Apr. 21, 2018, pp. 1–14, ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3173986. URL: https://dl.acm.org/doi/10.1145/3173574.3173986 (visited on 04/26/2024).

[47] A. Glaese, N. McAleese, M. Trębacz, J. Aslanides, V. Firoiu, T. Ewalds, M. Rauh, L. Weidinger, M. Chadwick, P. Thacker, *et al.*, "Improving alignment of dialogue agents via targeted human judgements," *arXiv preprint arXiv:2209.14375*, 2022.

[48] R. Liu, G. Zhang, X. Feng, and S. Vosoughi, "Aligning generative language models with human values," in *Findings of the Association for Computational Linguistics: NAACL 2022*, 2022, pp. 241–252.

[49] R. Liu, C. Jia, G. Zhang, Z. Zhuang, T. Liu, and S. Vosoughi, "Second thoughts are best: Learning to re-align with human values from text edits," *Advances in Neural Information Processing Systems*, vol. 35, pp. 181–196, 2022.

[50] S. Jung, G. Han, D. W. Nam, and K.-W. On, "Binary classifier optimization for large language model alignment," *arXiv preprint arXiv:2404.04656*, 2024.

[51] I. Solaiman and C. Dennison, "Process for adapting language models to society (palms) with values-targeted datasets," *Advances in Neural Information Processing Systems*, vol. 34, pp. 5861–5873, 2021.

[52] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, *et al.*, "Constitutional ai: Harmlessness from ai feedback," *arXiv preprint arXiv:2212.08073*, vol. 109, no. 1, pp. 25–42, 2022.

[53] Y. Dong, Z. Wang, M. N. Sreedhar, X. Wu, and O. Kuchaiev, "Steerlm: Attribute conditioned sft as an (user-steerable) alternative to rlhf," *arXiv preprint arXiv:2310.05344*, 2023.

[54] K. Yang and D. Klein, "Fudge: Controlled text generation with future discriminators," *arXiv preprint arXiv:2104.05218*, 2021.

[55] H. Wang, Y. Lin, W. Xiong, R. Yang, S. Diao, S. Qiu, H. Zhao, and T. Zhang, "Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards," *arXiv preprint arXiv:2402.18571*, 2024.

[56] A. Birhane, W. Isaac, V. Prabhakaran, M. Diaz, M. C. Elish, I. Gabriel, and S. 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*, 2022, pp. 1–8.

[57] F. Delgado, S. Yang, M. Madaio, and Q. Yang, "The participatory turn in ai design: Theoretical foundations and the current state of practice," in *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 2023, pp. 1–23.

[58] P. M. Asaro, "Transforming society by transforming technology: The science and politics of participatory design," *Accounting, Management and Information Technologies*, vol. 10, no. 4, pp. 257–290, 2000, ISSN: 0959-8022. DOI: https://doi.org/10.1016/S0959-8022(00)00004-7. URL: https://www.sciencedirect.com/science/article/pii/S0959802200000047.

[59] M. Sloane, E. Moss, O. Awomolo, and L. Forlano, *Participation is not a design fix for machine learning*, 2020. arXiv: 2007.02423 [cs.CY].

[60] A. F. Cooper, E. Moss, B. Laufer, and H. Nissenbaum, "Accountability in an algorithmic society: Relationality, responsibility, and robustness in machine learning," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 864–876.

[61] M. Sloane, "Controversies, contradiction, and "participation" in ai," *Big Data & Society*, vol. 11, no. 1, p. 20 539 517 241 235 862, 2024.

[62] J. Dzieza, *Inside the ai factory*, [Accessed 19-05-2024], 2023. URL: https://www.theverge.com/features/23764584/ai-artificial-intelligence-data-notation-labor-scale-surge-remotasks-openai-chatbots.

[63] H. Nissenbaum, "Accountability in a computerized society," *Science and engineering ethics*, vol. 2, pp. 25–42, 1996.

[64] K. Hao, "Artificial intelligence is creating a new colonial world order," *MIT Technology Review*, 2022.

[65] W. Nekoto, V. Marivate, T. Matsila, T. Fasubaa, T. Kolawole, T. Fagbohungbe, S. O. Akinola, S. H. Muhammad, S. Kabongo, S. Osei, *et al.*, "Participatory research for low-resourced machine translation: A case study in african languages," *arXiv preprint arXiv:2010.02353*, 2020.

[66] M. Feffer, M. Skirpan, Z. Lipton, and H. Heidari, "From preference elicitation to participatory ml: A critical survey & guidelines for future research," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 2023, pp. 38–48.

[67] H. Suresh, R. Movva, A. L. Dogan, R. Bhargava, I. Cruxên, Á. M. Cuba, G. Taurino, W. So, and C. D'Ignazio, "Towards intersectional feminist and participatory ml: A case study in supporting feminicide counterdata collection," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 667–678.

[68] M. K. Lee, D. Kusbit, A. Kahng, J. T. Kim, X. Yuan, A. Chan, D. See, R. Noothigattu, S. Lee, A. Psomas, *et al.*, "Webuildai: Participatory framework for algorithmic governance," *Proceedings of the ACM on human-computer interaction*, vol. 3, no. CSCW, pp. 1–35, 2019.

[69] L. Weidinger, K. R. McKee, R. Everett, S. Huang, T. O. Zhu, M. J. Chadwick, C. Summerfield, and I. Gabriel, "Using the veil of ignorance to align ai systems with principles of justice," *Proceedings of the National Academy of Sciences*, vol. 120, no. 18, e2213709120, 2023.

[70] B. Zhang and A. Dafoe, "Artificial intelligence: American attitudes and trends," *Available at SSRN 3312874*, 2019.

[71] S. Cave, K. Coughlan, and K. Dihal, "" scary robots" examining public responses to ai," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 331–337.

[72] B. Balaram, T. Greenham, and J. Leonard, "Artificial intelligence: Real public engagement," *RSA, London. Retrieved November*, vol. 5, p. 2018, 2018.

[73] T. A. T. I. Ada Lovelace Institute, *How do people feel about ai?* [Accessed 19-05-2024], 2023. URL: https://www.adalovelaceinstitute.org/wp-content/uploads/2023/06/Ada-Lovelace-Institute-The-Alan-Turing-Institute-How-do-people-feel-about-AI.pdf.

[74] A. Bewersdorff, X. Zhai, J. Roberts, and C. Nerdel, "Myths, mis- and preconceptions of artificial intelligence: A review of the literature," *Computers and Education: Artificial Intelligence*, vol. 4, p. 100 143, 2023, ISSN: 2666-920X. DOI: 10.1016/j.caeai.2023.100143. URL: https://www.sciencedirect.com/science/article/pii/S2666920X2300022X.

[75] S. Cave, C. Craig, K. Dihal, S. Dillon, J. Montgomery, B. Singler, and L. Taylor, "Portrayals and perceptions of ai and why they matter," 2018.

[76] B. Richardson, D. Prioleau, K. Alikhademi, and J. E. Gilbert, "Public accountability: Understanding sentiments towards artificial intelligence across dispositional identi-

ties," in *2020 IEEE International Symposium on Technology and Society (ISTAS)*, 2020, pp. 489–496. DOI: 10.1109/ISTAS50296.2020.9462184.

[77] H. B. Smith K., *Royalsociety.org*, [Accessed 19-05-2024], 2019. URL: https://royalsociety.org/-/media/policy/projects/science-education-tracker/5-science-education-tracker-2019-machine-learning.pdf.

[78] A. T. Brian Kennedy and E. Saks, *Public Awareness of Artificial Intelligence in Everyday Activities — pewresearch.org*, [Accessed 19-05-2024], 2023. URL: https://www.pewresearch.org/science/2023/02/15/public-awareness-of-artificial-intelligence-in-everyday-activities/.

[79] C. B. Leggon, "The impact of science and technology on african americans," *Humboldt Journal of Social Relations*, pp. 35–53, 1995.

[80] S. S. Smith, "Race and trust," *Annual Review of Sociology*, vol. 36, pp. 453–475, 2010.

[81] K. Akash, W.-L. Hu, T. Reid, and N. Jain, "Dynamic modeling of trust in human-machine interactions," in *2017 American Control Conference (ACC)*, IEEE, 2017, pp. 1542–1548.

[82] D. Gallimore, J. B. Lyons, T. Vo, S. Mahoney, and K. T. Wynne, "Trusting robocop: Gender-based effects on trust of an autonomous robot," *Frontiers in Psychology*, vol. 10, p. 482, 2019.

[83] H. Suresh, S. R. Gomez, K. K. Nam, and A. Satyanarayan, "Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–16.

[84] G. Adamson, "Explainable artificial intelligence (xai): A reason to believe?" *Law Context: A Socio-Legal J.*, vol. 37, p. 23, 2020.

[85] D. Long, J. Roberts, B. Magerko, K. Holstein, D. DiPaola, and F. Martin, "Ai literacy: Finding common threads between education, design, policy, and explainability," in *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '23, <conf-loc>, <city>Hamburg</city>, <coun-

try>Germany</country>, </conf-loc>: Association for Computing Machinery, 2023, ISBN: 9781450394222. DOI: 10.1145/3544549.3573808. URL: https://doi.org/10.1145/3544549.3573808.

[86] T. Kihara, R. Bendor, and D. Lomas, "Designing an escape room in the city for public engagement with ai-enhanced surveillance," ser. CHI EA '19, Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–6, ISBN: 9781450359719. DOI: 10.1145/3290607.3313003. URL: https://doi.org/10.1145/3290607.3313003.

[87] D. A. Robb, M. I. Ahmad, C. Tiseo, *et al.*, "Robots in the danger zone: Exploring public perception through engagement," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '20, ACM, Mar. 2020. DOI: 10.1145/3319502.3374789. URL: http://dx.doi.org/10.1145/3319502.3374789.

[88] E. Ballard, K. Werner, and P. Priyadarshini, "Boundary objects in translation: The role of language in participatory system dynamics modeling," *System Dynamics Review*, vol. 37, no. 4, pp. 310–332, 2021.

[89] C. D'Ignazio, "Creative data literacy: Bridging the gap between the data-haves and data-have nots," *Information Design Journal*, vol. 23, no. 1, pp. 6–18, 2017.

[90] R. Bhargava, R. Kadouaki, E. Bhargava, G. Castro, and C. D'Ignazio, "Data murals: Using the arts to build data literacy," *The Journal of Community Informatics*, vol. 12, no. 3, 2016.

[91] E. Sulmont, E. Patitsas, and J. R. Cooperstock, "Can you teach me to machine learn?" In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, ser. SIGCSE '19, Minneapolis, MN, USA: Association for Computing Machinery, 2019, pp. 948–954, ISBN: 9781450358903. DOI: 10.1145/3287324.3287392. URL: https://doi.org/10.1145/3287324.3287392.

[92] M. Katell, M. Young, B. Herman, D. Dailey, A. Tam, V. Guetler, C. Binz, D. Raz, and P. M. Krafft, *An algorithmic equity toolkit for technology audits by community advocates and activists*, 2019. arXiv: 1912.02943 [cs.CY].

[93] M. Bucholtz, A. Lopez, A. Mojarro, E. Skapoulli, C. VanderStouwe, and S. Warner-Garcia, "Sociolinguistic justice in the schools: Student researchers as linguistic experts," *Language and Linguistics Compass*, vol. 8, no. 4, pp. 144–157, 2014.

[94] C. Frauenberger, J. Good, G. Fitzpatrick, and O. S. Iversen, "In pursuit of rigour and accountability in participatory design," *International journal of human-computer studies*, vol. 74, pp. 93–106, 2015.

[95] R. Rorty, *Philosophy and the Mirror of Nature*. Princeton university press, 2009.

[96] O. S. Iversen, K. Halskov, and T. W. Leong, "Values-led participatory design," *CoDesign*, vol. 8, no. 2-3, pp. 87–103, 2012.

[97] M. Hammersley, "Emergent design," *The SAGE Handbook of Qualitative Research Design*, pp. 55–68, 2022.

[98] J. A. Maxwell, *Qualitative research design: An interactive approach*. Sage publications, 2012.

[99] K. Charmaz, *Constructing grounded theory: A practical guide through qualitative analysis*. sage, 2006.

[100] K. Charmaz, "The power of constructivist grounded theory for critical inquiry," *Qualitative inquiry*, vol. 23, no. 1, pp. 34–45, 2017.

[101] K. Charmaz, "Grounded theory: Objectivist and constructivist methods," *Handbook of qualitative research*, vol. 2, no. 1, pp. 509–535, 2000.

[102] Y. Chun Tie, M. Birks, and K. Francis, "Grounded theory research: A design framework for novice researchers," *SAGE open medicine*, vol. 7, p. 2 050 312 118 822 927, 2019.

[103] K. Holtzblatt and H. Beyer, *Contextual design: defining customer-centered systems*. Elsevier, 1997.

[104] D. Rodighiero, "Mapping affinities: Visualizing academic practice through collaboration," EPFL, Tech. Rep., 2018.

[105] Y. C. Tie, M. Birks, and K. Francis, "Grounded theory research: A design framework for novice researchers," *SAGE Open Medicine*, vol. 7, p. 2 050 312 118 822 927, 2019, PMID: 30637106. DOI: 10.1177/2050312118822927. eprint: https://doi.org/10.1177/2050312118822927. URL: https://doi.org/10.1177/2050312118822927.

[106] M. Price, "Open coding for machine learning," Ph.D. dissertation, Massachusetts Institute of Technology, 2022.

[107] Q. Chen, J. Bragg, L. B. Chilton, and D. S. Weld, "Cicero: Multi-turn, contextual argumentation for accurate crowdsourcing," in *Proceedings of the 2019 chi conference on human factors in computing systems*, 2019, pp. 1–14.

[108] C. Small, M. Bjorkegren, T. Erkkilä, L. Shaw, and C. Megill, "Polis: Scaling deliberation by mapping high dimensional opinion spaces," *Recerca: revista de pensament i anàlisi*, vol. 26, no. 2, 2021.

[109] J. J. Y. Chung, J. Y. Song, S. Kutty, S. Hong, J. Kim, and W. S. Lasecki, "Efficient elicitation approaches to estimate collective crowd answers," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–25, 2019.

[110] A. M. Davani, M. Díaz, and V. Prabhakaran, "Dealing with disagreements: Looking beyond the majority vote in subjective annotations," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 92–110, 2022.

[111] S. Fish, P. Gölz, D. C. Parkes, A. D. Procaccia, G. Rusak, I. Shapira, and M. Wüthrich, "Generative social choice," *arXiv preprint arXiv:2309.01291*, 2023.

[112] N. Sambasivan, E. Arnesen, B. Hutchinson, T. Doshi, and V. Prabhakaran, *Re-imagining algorithmic fairness in india and beyond*, 2021. arXiv: 2101.09995 [cs.CY].

[113] S. Bhatt, S. Dev, P. Talukdar, S. Dave, and V. Prabhakaran, *Re-contextualizing fairness in nlp: The case of india*, 2022. arXiv: 2209.12226 [cs.CL].

[114] S. Jasanoff, *The Ethics of Invention: Technology and the Human Future*. New York: W.W. Norton & Company, 2016.

[115] J. Thatcher, D. O'Sullivan, and D. Mahmoudi, "Data colonialism through accumulation by dispossession: New metaphors for daily data," *Environment and Planning D: Society and Space*, vol. 34, no. 6, pp. 990–1006, 2016. DOI: 10.1177/0263775816633195. eprint: https://doi.org/10.1177/0263775816633195.

[116] N. Couldry and U. A. Mejías, "Data colonialism: Rethinking big data's relation to the contemporary subject," *Television & New Media*, vol. 20, pp. 336–349, 2018. URL: https://api.semanticscholar.org/CorpusID:150041824.

[117] S. Milan and E. Treré, *Big Data from the South(s): Beyond Data Universalism*, (1) BigDatafromtheSouth(s):BeyondDataUniversalism.https://papers.ssrn.com/abstract=3384569(accessed2024-05-13)., [Accessed 13-05-2024].

[118] B. Green, ""good" isn't good enough," 2019. URL: https://api.semanticscholar.org/CorpusID:209379533.

[119] J. Sylvester and E. Raff, *What about applied fairness?* 2018. arXiv: 1806.05250 [cs.AI].

[120] *Pause Giant AI Experiments: An Open Letter - Future of Life Institute — futureoflife.org*, https://futureoflife.org/open-letter/pause-giant-ai-experiments/, [Accessed 13-05-2024].

[121] H. Oberdiek and M. Tiles, *Living in a Technological Culture: Human Tools and Human Values*, H. Oberdiek, Ed. New York: Routledge, 1995.

[122] L. Winner, "Do artifacts have politics?," pp. 177–192, 2017.

[123] B. Hutchinson, N. Rostamzadeh, C. Greer, K. Heller, and V. Prabhakaran, *Evaluation gaps in machine learning practice*, 2022. arXiv: 2205.05256 [cs.LG].

[124] S. R. Arnstein, "A ladder of citizen participation," *Journal of the American Institute of Planners*, vol. 35, no. 4, pp. 216–224, 1969. DOI: 10.1080/01944366908977225. eprint: https://doi.org/10.1080/01944366908977225. URL: https://doi.org/10.1080/01944366908977225.

[125] I. Rahwan, "Society-in-the-loop: Programming the algorithmic social contract," *Ethics and information technology*, vol. 20, no. 1, pp. 5–14, 2018.

[126] Y. Wang, S. R. Nagireddy, C. T. Thota, D. H. Ho, and Y. Lee, "Community-in-the-loop: Creating artificial process intelligence for co-production of city service," *Proc. ACM Hum.-Comput. Interact.*, vol. 6, no. CSCW2, Nov. 2022. DOI: 10.1145/3555176. URL: https://doi.org/10.1145/3555176.

[127] V. Prabhakaran and D. Martin Jr, "Participatory machine learning using community-based system dynamics," *Health and Human Rights*, vol. 22, no. 2, p. 71, 2020.

[128] A. Clement and P. Van den Besselaar, "A retrospective look at pd projects," *Commun. ACM*, vol. 36, no. 6, pp. 29–37, Jun. 1993, ISSN: 0001-0782. DOI: 10.1145/153571.163264. URL: https://doi.org/10.1145/153571.163264.

[129] T. Bratteteig and I. Wagner, "Disentangling power and decision-making in participatory design," in *Proceedings of the 12th Participatory Design Conference: Research Papers - Volume 1*, ser. PDC '12, Roskilde, Denmark: Association for Computing Machinery, 2012, pp. 41–50, ISBN: 9781450308465. DOI: 10.1145/2347635.2347642. URL: https://doi.org/10.1145/2347635.2347642.

[130] J. A. Kuhlberg, I. Headen, E. A. Ballard, and D. M. J. au2, *Advancing community engaged approaches to identifying structural drivers of racial bias in health diagnostic algorithms*, 2023. arXiv: 2305.13485 [cs.LG].

[131] S. Costanza-Chock, *Design justice: Community-led practices to build the worlds we need*. Cambridge, MA: The MIT Press, 2020.

[132] S. Girish and M. Avery, "Data cooperative: Enabling meaningful collective negotiation of data rights for communities," *Available at SSRN 4414473*, 2022, [Accessed 19-05-2024]. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4414473.

[133] A. Institute, *He journey of enyorata loviluku women's group: A data co-op in action*, [Accessed 19-05-2024], 2024. URL: https://aapti.in/blog/the-journey-of-enyorata-loviluku-womens-group-a-data-co-op-in-action/.

[134] S. Manohar, *Trust law, fiduciaries,and data trusts*, [Accessed 19-05-2024], 2020. URL: https://thedataeconomylab.com/wp-content/uploads/2020/10/DataTrustsPpr_SM.pdf.

[135] "INDIGENOUS AI," INDIGENOUS AI. (), URL: https://www.indigenous-ai.net/abundant/ (visited on 05/15/2024).

[136] I. D. of Indigenous Languages, *Leveraging unesco normative instruments for an ethical generative ai use of indigenous data*, [Accessed 19-05-2024], 2023. URL: https://www.unesco.org/en/articles/leveraging-unesco-normative-instruments-ethical-generative-ai-use-indigenous-data.

[137] "Blending Indigenous Knowledge and artificial intelligence to enable adaptation," WWF Arctic. (), URL: https://www.arcticwwf.org/the-circle/stories/blending-indigenous-knowledge-and-artificial-intelligence-to-enable-adaptation/ (visited on 05/15/2024).

[138] J. E. Lewis, A. Abdilla, N. Arista, *et al.*, "Indigenous protocol and artificial intelligence position paper," Aboriginal Territories in Cyberspace, Honolulu, HI, Project Report 10.11573/spectrum.library.concordia.ca.00986506, 2020, Edited by Jason Edward Lewis. English Language Version of "Ka?ina Hana ?Ōiwi a me ka Waihona ?Ike Hakuhia Pepa Kūlana" available at: https://spectrum.library.concordia.ca/id/eprint/990094/. URL: https://spectrum.library.concordia.ca/id/eprint/986506/.

[139] O. O. Queerinai, A. Ovalle, A. Subramonian, A. Singh, C. Voelcker, D. J. Sutherland, D. Locatelli, E. Breznik, F. Klubicka, H. Yuan, *et al.*, "Queer in ai: A case study in community-led participatory ai," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 1882–1895.

[140] D. J. Fourie, "Mechanisms to improve citizen participation in government and its administration," *South African Journal of Economic and Management Sciences*, vol. 4, no. 1, pp. 216–233, 2001.

[141] PCAST, *Pcast releases letter on advancing public engagement with the sciences*, [Accessed 19-05-2024], 2023. URL: https://www.whitehouse.gov/wp-content/uploads/2023/08/PCAST_Science-Engagement-Letter_August2023.pdf.

[142] H. Riesch and C. Potter, "Citizen science as seen by scientists: Methodological, epistemological and ethical dimensions," *Public understanding of science*, vol. 23, no. 1, pp. 107–120, 2014.

[143] J. P. Woolley, M. L. McGowan, H. J. Teare, V. Coathup, J. R. Fishman, R. A. Settersten, S. Sterckx, J. Kaye, and E. T. Juengst, "Citizen science or scientific citizenship? disentangling the uses of public engagement rhetoric in national research initiatives," *BMC medical ethics*, vol. 17, no. 1, pp. 1–17, 2016.

[144] I. Shapiro, "Enough of deliberation: Politics is about interests and power," 1999.

[145] W. E. Connolly, "Legitimacy and the state," *(No Title)*, 1984.

[146] J. Parkinson, "Hearing voices: Negotiating representation claims in public deliberation," *The British Journal of Politics and International Relations*, vol. 6, no. 3, pp. 370–388, 2004.

[147] P. Jones and, C. Lorne, C. Speed, C. Durose, and L. Richardson, "Using technology to help communities shout louder," in *Designing Public Policy for Co-production: Theory, Practice and Change*. Bristol University Press, 2015, pp. 141–148.

[148] J. Jull, A. Giles, and I. D. Graham, "Community-based participatory research and integrated knowledge translation: Advancing the co-creation of knowledge," *Implementation science*, vol. 12, pp. 1–9, 2017.

[149] A. M. Fejerskov, "The new technopolitics of development and the global south as a laboratory of technological experimentation," *Science, Technology, & Human Values*, vol. 42, no. 5, pp. 947–968, 2017. DOI: 10.1177/0162243917709934. eprint: https://doi.org/10.1177/0162243917709934. URL: https://doi.org/10.1177/0162243917709934.