Community-Centered AI is Feasible and Beneficial for Impacted Communities

Tzu-Sheng Kuo
Carnegie Mellon University
Pittsburgh, PA, USA
tzushenk@cs.cmu.edu

Quan Ze Chen
AI & Democracy Foundation
Seattle, WA, USA
jim@ai-democracy.org

Amy X. Zhang
University of Washington
Seattle, WA, USA
axz@cs.uw.edu

Kenneth Holstein*
EPFL
Lausanne, Switzerland
ken.holstein@epfl.ch

Haiyi Zhu*
Carnegie Mellon University
Pittsburgh, PA, USA
haiyiz@cs.cmu.edu

Abstract

AI systems are increasingly used in community contexts, yet they often fail to align with the communities they impact. To address this misalignment, we advocate for community-centered AI—an AI development paradigm that shifts away from simply pursuing general-purpose AI and instead focuses on co-developing tailored AI systems with the specific communities they impact. This position paper argues that community-centered AI is both feasible and beneficial for impacted communities. To support this argument, we first outline benefits of community-centered AI, including greater alignment, enhanced legitimacy, capacity building, and normative justification. Next, we highlight emerging efforts in community-centered AI across the AI development process, demonstrating its feasibility in delivering these benefits. Finally, we propose a framework to guide future work in this area, characterizing community-centered AI efforts along three dimensions: participation mode, decision-making process, and development layer. Overall, this paper calls for increased research on community-centered AI.

1 Introduction

Human lives are fundamentally shaped by the *communities* to which we belong [150]. In this paper, we adopt the sociological notion of community, which broadly refers to groups of people connected by demographic, socioeconomic, cultural, political, and affective ties that foster a strong sense of collective identity [131]. Examples of communities include residents of a neighborhood, users of a specific subreddit, or members of an Indigenous Māori tribe [54]. Whether online or offline, small or large, each of these communities is bound by shared identities, values, and norms [14, 81].

Today, AI is increasingly used in community contexts, transforming how communities function. For example, local governments have employed AI to allocate social services [9], such as public housing [87, 108] and child welfare [19, 73, 129]. Online communities have similarly used AI to assist or automate tasks such as content moderation [132], including subreddits [65], Discord servers [86], and different language editions of Wikipedia [50, 138]. AI deployments have also been shown to have disparate impacts on identity-based communities [20], such as Black or Queer communities [15, 120]. In each of these examples, AI serves as a form of digital *community infrastructure* [103], fundamentally shaping people's experiences and interactions within a given community.

*Co-senior authors contributed equally.

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Position Paper Track.

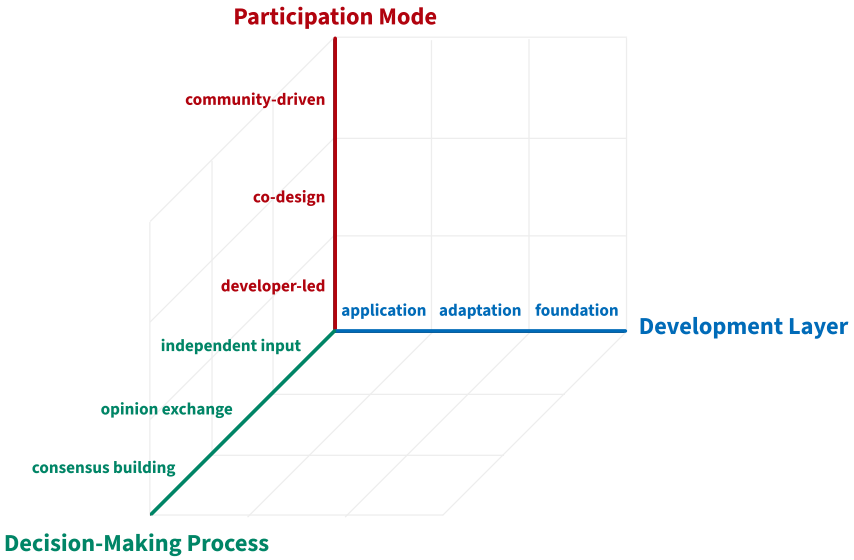


Figure 1: The community-centered AI framework. It has three dimensions: 1) participation mode, which concerns who drives the AI development process; 2) decision-making process, which describes how members of a community reach collective decisions that shape AI development; and 3) development layer, which specifies where in the AI stack the work takes place. These three dimensions constitute a space for understanding design variations in community-centered AI.

However, as AI systems move from development labs to real-world communities, they often fail to align with the communities they impact [2, 7, 32, 74]. Prior research has reported various cases in which AI deployments have clashed with the actual needs and values of real-world communities, whether in neighborhoods [24, 126, 148], online spaces [51], or identity-based groups [128, 169]. A primary cause is that developers often lack the time, incentives, and contextual knowledge to align AI systems with individual communities [158, 177]. Meanwhile, community members have expressed a strong desire to participate in the development of AI systems that impact their lives, if given the opportunity [28, 140, 145].

In this paper, we advocate for **community-centered AI**—an AI development paradigm that shifts away from simply pursuing general-purpose AI and instead focuses on co-developing tailored AI systems with the specific communities they impact. This argument is timely and necessary: while AI is rapidly deployed in community contexts, ML research remains largely focused on generalization across individual users (micro-level) or on AI's broad societal impacts (macro-level), leaving community-centered AI (meso-level) neglected [28, 44, 71, 147]. Prior critiques attribute this neglect to the ML field's pursuit of universality and scale [11, 123], under which tailoring AI to local communities is often dismissed as too narrow to be valuable or too idealistic to succeed.

This position paper argues that community-centered AI is both feasible and beneficial for impacted communities.[2] We first outline four potential benefits: greater alignment, enhanced legitimacy, capacity building, and normative justification (§2). To demonstrate the feasibility of realizing these benefits, we present emerging efforts in community-centered AI across the AI development process (§3). We then introduce the **community-centered AI framework**, as shown in Figure 1, which characterizes efforts in this space along three dimensions: participation mode, decision-making process, and development layer (§4), with the goal of motivating future work (§5). Finally, we discuss counterarguments (§6) and related concepts like human-centered AI and participatory AI (§7).

[2]It is important to clarify that we are not suggesting *all* AI development should be community-centered. However, for AI designed for use in community contexts, a community-centered approach is beneficial, if not necessary.

2 Potential Benefits of Community-Centered AI

2.1 Greater Alignment

AI alignment refers to the effort to ensure that AI systems are aligned with human values and goals [66]. In the context of community-centered AI, the reference point for alignment is not random crowdworkers or an unspecified public, but rather a specific community in which the AI system is used [7]. Therefore, instead of optimizing AI alignment based on standardized benchmarks that are often framed as universally applicable [123], AI alignment in community contexts focuses on the preferences, satisfaction, and behavioral outcomes of the specific impacted community [7, 37, 176]. We use the term alignment broadly to encompass how AI is developed, used, and governed in community contexts, rather than restricting it to the training of models in isolation.

Aligning AI with individual communities is necessary because human society is pluralistic [164], with each community possessing distinct norms and priorities. For example, online communities have different expectations for acceptable content on their platforms [38, 93], and different cultures hold varying standards of fairness for distributing shared resources [106, 127]. Thus, when AI systems are deployed in community contexts for tasks such as content moderation and resource allocation, communities also hold divergent expectations for how these AI systems should behave [16, 50].

Community-centered AI offers an epistemic advantage for AI alignment because community members are experts in their own contexts. As standpoint theory suggests [21, 57], this position equips them with situated knowledge that outsiders often lack [56]. Their participation may help better identify alignment targets and provide input to steer AI towards these targets, thereby achieving greater alignment. Evidence from software development [22, 79, 104], urban planning [94], and healthcare design [152, 162] shows that community participation leads to outcomes better aligned with local contexts. AI development can similarly benefit from community-centered processes.

2.2 Enhanced Legitimacy

Legitimacy is a subject of extensive research in sociology [68], political science [99], and philosophy [49], and broadly refers to the perception that an entity's actions are appropriate within a socially constructed system of norms and values [146]. For AI deployed in community contexts, legitimacy refers to whether AI is perceived as appropriate by the community it impacts [163]. Beyond outcome-level alignment, legitimacy (or lack thereof) can stem from whether the community gets to shape the *process* of AI development and governance [84, 92], a concept known as procedural legitimacy [151].

AI systems and policies can face a legitimacy crisis when introduced in communities without their involvement, leading to pushback or abandonment [74, 163]. For example, volunteer moderators on Stack Overflow organized a strike after the platform enforced a new policy on AI-generated content without consulting the community [167]. Rather than opposing having an AI policy, moderators protested the exclusionary process that ignored their role in governance. The strike ended only after the platform worked with moderators to co-author a revised policy [167].

Community-centered AI can potentially address AI's legitimacy crisis by involving communities in the AI development and governance process. This potential is grounded in research on civic engagement, where public participation increases the perceived legitimacy of policy-making by giving people a voice, signaling respect, and making difficult trade-offs transparent [33, 116, 151]. In the context of AI, such participation may similarly increase acceptance [92].

2.3 Capacity Building

The United Nations defines capacity building as the process of developing and strengthening the skills, instincts, and abilities that communities require to survive, adapt, and thrive [153]. Capacity-building initiatives in other fields include providing digital literacy training to local small business owners to support economic mobility [64, 79], or training local health workers to establish sustainable public health expertise [96]. These initiatives focus on long-term empowerment by building skills and knowledge, rather than offering immediate solutions that may not be sustainable.

Community-centered AI has the potential to enhance capacity building by improving communities' AI literacy [101]. By actively involving communities in the AI development process, members may gain a better understanding of AI models' capabilities and limitations. This kind of AI literacy is

critical, similar to educational efforts that aim to equip students with the skills needed to use AI effectively and critically [59, 156]. Without sufficient AI literacy, there is a risk of misusing AI, such as over-reliance on these systems or underutilization of their potential [141].

Community-centered AI also has the potential to enhance capacity building by offering communities opportunities for collective ideation, deliberation, and reflection on their needs and desires related to AI. This process of collective sensemaking is critical, if not necessary, as communities are not monolithic and individual members may hold diverse perspectives [46, 47, 142]. The opportunity for community members to exchange perspectives may promote mutual understanding, foster internal learning, and facilitate consensus-building in AI development and governance [34].

2.4 Normative Justification

Finally, community-centered AI aligns with the normative principle of self-governance [27]. As Ostrom stated in her influential principles for governing the commons [116]: "*individuals affected by the operational rules should be able to participate in modifying those rules.*" As AI systems increasingly serve as digital infrastructure that enacts the rules of community governance [52], such as distributing common resources and mediating member interactions, it becomes a normative ideal that communities impacted by AI should have the opportunity to participate in shaping these systems.

3 Emerging Examples of Community-Centered AI

To illustrate the feasibility of community-centered AI, this section presents emerging examples and connects their reported benefits to those discussed above (§2). These examples span the AI development pipeline, including problem formulation, data curation, training, evaluation, and governance. Covering the entire pipeline is crucial because community-centered AI is a holistic development approach. Failures to align AI systems with the communities they impact at early stages, such as problem formulation, cannot be resolved through downstream model training efforts alone [69].

3.1 Problem Formulation

Problem formulation is the first step in AI development [119], determining whether and how an AI system should be implemented in a given context [170]. Early decisions made during problem formulation can impact subsequent stages of AI development [171], and can be difficult to remediate retrospectively [19, 75, 148, 163]. For example, a predictive policing tool launched in Pittsburgh faced widespread protests and was later banned by the city council because residents fundamentally rejected the premise of deploying predictive AI for policing [69].

To ensure AI systems are designed with goals that reflect community needs from the outset, researchers have increasingly involved impacted communities early in the problem formulation stage. For example, in the context of AI used within the U.S. child welfare system [129], researchers conducted design workshops with social workers and impacted parents to reimagine how AI could address the challenges they observe based on their lived experiences [145]. When given the opportunity to help re-formulate the problem, participants articulated solutions that more closely aligned with their needs and practices (§2.1), including no-AI or no-tech alternatives that should be accepted as valid outcomes [42]. To create design materials reusable across a wider range of contexts, other work has developed design toolkits such as the Situate AI Guidebook [72] and AI Lifecycle Comicboarding [87], which include deliberation questions and open-ended comic boards to elicit and encourage discussion among community stakeholders about their desiderata for AI systems used in their communities. Research has found that such approaches can help community members meaningfully critique consequential "technical" decisions behind an AI system design, and pose alternative design directions, even without prior AI literacy [55, 87, 115] (§2.1).

3.2 Data Curation

Data curation is an intentional process that involves selecting what data to include in a dataset and, in the case of labeled datasets, deciding how each data point should be labeled [109, 110, 113]. Because standardized AI datasets often fail to capture the specific context of local communities [30, 123, 128], research has advocated drawing on lessons from community archives and libraries

[67], where directly impacted communities contribute data intended to represent them [8]. For example, the Local Intersectional Visual Spaces (LIVS) dataset [112] was developed in collaboration with 30 community organizations in Montreal and was used to fine-tune a text-to-image model via Direct Preference Optimization [122] to generate representations of the city. This research showed that the model fine-tuned with the LIVS dataset is more aligned with local residents' preferences for public space design, compared to models not tailored to the local context (§2.1). In the same vein, Wikibench [85] is a system that enables Wikipedians from specific language communities, such as the English or Dutch Wikipedia, to collaboratively curate datasets and author datasheets for AI content moderation tools used on Wikipedia in their respective languages [50]. This research demonstrated that collaboratively curated datasets are more closely aligned with their moderation norms than those labeled by individuals alone [85] (§2.1). This collaborative process also encouraged community members to reflect on their moderation practices, contributing to capacity building within the community [52] (§2.3).

3.3 Training

As AI models have advanced from traditional ML classifiers [82, 91, 136] to generative AI [121, 157], model training can refer to a range of methods, including training models from scratch [13], fine-tuning pretrained models [31, 60, 117, 122], and designing system prompts for language models [35, 173].

Emerging research has recognized the limitations of generic models and sought to incorporate community input into the training process to develop models better tailored to specific communities. For traditional ML models, WeBuildAI [92] is a framework that facilitates the development of a community-based food donation matching system shaped by input from local stakeholders, including donors, volunteers, and recipient organizations, who found the distribution procedures better aligned with their local notions of fairness (§2.2). For LLMs, unlike post-training techniques like RLHF that aggregate contextless preferences from random crowdworkers [117], Collective Constitutional AI [61] gathers statements from U.S. population samples to create a constitution for fine-tuning language models. This approach has been shown to produce outputs that more closely align with U.S. public values, compared to models fine-tuned using a constitution developed by developers alone [5] (§2.1). Relatedly, Modular Pluralism [37], an example of pluralistic alignment [142], plugs smaller, specialized language models into a base LLM to better align with the preferences of individual communities (§2.1). For system prompt design, systems such as Botender [86] enable online communities to collaboratively design the system prompt for the LLM that powers their own AI bots. Users found that bot behavior shaped by prompts they designed themselves was better aligned with their unique needs and norms [86] (§2.1). Moreover, the collaborative design process fosters reflection and discussion about the desirable bot behavior within the community [86] (§2.3).

3.4 Evaluation

Evaluation determines whether a model's behaviors effectively address community needs and adhere to local norms. This includes a range of activities, such as auditing AI failures [105], constructing measurements [48, 75, 161], and developing benchmarks for systematic evaluation [85, 123].

Most efforts in this area have focused on AI auditing [29, 80, 89, 139], with emerging work supporting the design of community-specific measurements and benchmarks. In the context of AI auditing, for example, researchers have collaborated with 22 public school educators in Hawai'i, where a statewide mandate requires the integration of Hawaiian language and culture into education, to audit the risks of using generative AI in classrooms, such as the misrepresentation of local culture [175]. Beyond auditing, recent work engaged residents of Kerala, residents of Tamil Nadu, and blind and low-vision individuals in the UK to co-construct measurement instruments for cultural appropriateness in text-to-image models [70]. To develop community-centered benchmarks, the Samiksha evaluation pipeline was co-created with civil society organizations and everyday users in India to assess a healthcare chatbot that operates in Indian languages [53]. These efforts reveal issues that require local knowledge for improved alignment [7] (§2.1), and enable community members to better understand the strengths and limitations of AI, thereby fostering AI literacy within the community [95] (§2.3).

3.5 Governance

AI governance refers to the processes and structures through which decisions related to AI are made, implemented, and enforced [124]. Besides existing efforts in technical AI governance [125], such as data regulation and compute monitoring, recent work advocates a sociotechnical approach [36] that accounts for societal factors, such as organizational bureaucracy, human labor, and power [17, 83].

Most efforts in AI governance focus on national and supranational initiatives [135], such as the White House Executive Order on AI [114] and the EU AI Act [159]. While a nation may be considered a broad community [12], recent research has increasingly focused on engaging local communities in AI governance within more specific contexts [26, 88]. For example, researchers in the UK worked with the local Muslim community to identify desirable Islamic ethical values for AI that are not represented in the UK AI White Paper, such as Shūrā, an Islamic practice of consultation that requires leaders to engage in collective decision-making with the community [16] (§2.4). To formalize the process of engaging community input in shaping AI governance policies, recent work has investigated how documentation practices [166], public procurement mechanisms [63], and digital tools such as PolicyCraft [84] can facilitate this engagement and lead to policies that are better supported by the impacted communities [84] (§2.2). For online communities, researchers have also examined how individual communities, such as subreddits [100] or other decentralized social media platforms [174], govern AI-generated content by developing policies that are aligned with their norms [43] (§2.1).

4 The Community-Centered AI Framework

To further elucidate the different forms of community-centered AI discussed previously (§3), we introduce a framework that characterizes community-centered AI along three dimensions: participation mode, decision-making process, and development layer, as illustrated in Figure 1. These three dimensions constitute a space in which the varieties of community-centered AI can be situated. Our framework is inspired by the Democracy Cube [41] in political science, which maps different forms of civic participation in democratic governance across a three-dimensional space. We distill relevant dimensions from the Democracy Cube and integrate supplementary taxonomies to tailor our framework to community-centered AI. Similar to the Democracy Cube, our goal is to help understand design variations in community-centered AI, rather than developing a general theory or prescribing a single best approach. This section describes each dimension in detail using examples from the previous section (§3), and discusses how the framework can be used in practice.

4.1 Participation Mode

The participation mode concerns who drives the AI development process. The categories within this dimension are informed by prior scholarship, including Arnstein's Ladder of Civic Participation [4] and the Parameters of Participation in AI [28], and reflect the degree of power and ownership communities have in making decisions about AI development [10, 22, 23, 97].

The participation mode encompasses three main categories: (1) development led by AI developers; (2) co-design between community stakeholders and AI developers; and (3) initiatives primarily driven by the community itself. An example of a developer-led effort is Collective Constitutional AI [61] (§3.3), where developers select eligible participants, decide how to process community input, and oversee model training and evaluation. In this approach, developers retain control throughout the AI development process, with the community primarily providing input for alignment. Co-design efforts such as the LIVS dataset project [112] (§3.2) conduct participatory action research [58], where researchers engage local community members as co-inquirers to collaboratively frame research questions and establish dataset inclusion criteria. In contrast, Wikibench [85] (§3.2) exemplifies a community-driven approach, empowering community members themselves to collaboratively build and curate their own datasets.

4.2 Decision-Making Process

This dimension examines how members of a community reach collective decisions that shape AI development, recognizing that even within a single community, members may hold diverse perspectives [62]. The categories are informed by decision-making models such as deliberative democracy [39, 149], which differentiate between mere voting and more participatory processes.

The decision-making processes are broadly classified into three categories: (1) independent community input without communication; (2) opinion exchange involving collective sensemaking, such as discussion; and (3) consensus-building, where the community actively makes collective decisions based on consensus. For community-centered AI, examples of independent input often involve recruiting participants with specific cultural backgrounds. For instance, in the Samiksha evaluation pipeline [53] (§3.4), Indian data workers are recruited to independently contribute chatbot queries in local languages, drawing on their lived experiences. Recognizing that individual perspectives may evolve when exposed to those of others, initiatives such as the Situate AI Guidebook [72] (§3.1) aim to bring community stakeholders together to interactively deliberate on their views, moving beyond unidirectional input to developers. To further facilitate consensus-building beyond the exchange of opinions, researchers have developed tools such as PolicyCraft [84] (§3.5), which enable community members to collaboratively draft AI governance policies in a Wikipedia style [77], allowing them to iteratively propose, critique, and refine AI policies collectively.

4.3 Development Layer

The development layers identify where in the AI stack community-centered AI efforts take place. The categories are informed by the three layers outlined in a blueprint for participatory foundation models [147], although we find that these layers are also applicable to traditional ML models. These layers reflect the extent to which community members can tailor models to their specific contexts.

The development layers consist of three levels: (1) the application layer, which involves directly adopting a pre-trained model, either a traditional ML model or a foundation model with prompt-level adjustments; (2) the adaptation layer, which refers to modifying an existing model through techniques such as transfer learning [107, 172], fine-tuning [112], RLHF [117], or other alignment mechanisms [46, 142]; and (3) the foundation layer, which involves building a model from scratch specifically for a community context. An example at the application layer is Botender [86] (§3.3), which enables online communities to collaboratively write system prompts for their AI agents. Examples at the adaptation layer include Modular Pluralism [37] (§3.3), where community input is used to fine-tune a specialized LLM for the specific needs of individual communities. The foundation layer is less commonly used for LLMs due to the costs of pre-training [147], but is more prevalent in traditional ML with smaller models. For instance, WeBuildAI enables community stakeholders to develop a food donation matching model that reflects their values in balancing equity and efficiency [92] (§3.3).

4.4 Immediate Use of the Framework

The community-centered AI framework can be used immediately in practice in at least two ways: as a *diagnostic tool* for AI developers, and as a *negotiation tool* for impacted communities. For AI developers pursuing community-centered AI, the framework supports critical reflection on current practices and surfaces alternative forms of community engagement that might otherwise be overlooked. For example, if a developer finds that their evaluation pipeline is purely developer-led and relies solely on independent, individual input, they may consider moving one notch along the axes, such as toward co-design or opinion exchange, to grant the community greater agency. Meanwhile, for impacted communities that desire more power-sharing in AI development or governance, the framework can serve as a concrete reference to advocate for alternative modes of engagement with AI developers.

5 Future Research

Beyond the immediate use of our framework (§4.4), this section outlines longer-term directions for advancing community-centered AI from two perspectives: gaps and opportunities surfaced by our framework (§5.1), and future agendas for researchers and policymakers (§5.2).

5.1 Gaps and Opportunities in the Framework

Beyond characterizing existing efforts, our framework surfaces three directions for future research. First, existing efforts are primarily clustered near the intersection of the three axes in the framework (Figure 1). In contrast, the opposite corner, which represents community-driven initiatives using consensus-building processes and foundational-level customization, remains unexplored. This is not to suggest that all community-centered AI should move toward that corner. Rather, it underscores a

current lack of tools and methods for communities that seek greater autonomy in developing their AI models. Second, increased participation may not monotonically result in greater benefits. Future research should thus investigate at which stages of the AI development process (§3), and through which specific forms of community engagement (§4), participation is most effective for achieving particular benefits (§2). This would enable communities to strategically focus their efforts, rather than wasting resources without meaningful gains. Finally, the current three dimensions of the framework, along with their corresponding categories, are not intended to be exhaustive. Future research could identify additional dimensions and categories that warrant attention, helping researchers better understand design trade-offs and recognize underexplored areas within the field.

5.2 Future Agendas for Researchers and Policymakers

For AI researchers, community-centered AI opens new directions for developing ML algorithms. For example, today's algorithms often rely on explicit human demonstrations or preference feedback for model alignment. However, in community contexts, rich traces of deliberation usually exist naturally, such as members' discussions on shared issues. This creates an opportunity to complement the current ML paradigms of *learning from demonstration* [130] and *learning from human feedback* [117] with what we term *learning from deliberation*. Directly leveraging these organic traces of deliberation for alignment could not only reduce the overhead of providing explicit labels, but also enable AI to capture latent community knowledge that is lost in isolated labels. Yet, it remains an open question how algorithms can distill valuable signals from noisy discussions while preserving community agency over what AI learns. Beyond this example, researchers can explore other alignment challenges unique to community contexts, such as modeling *power* as a parameter in pluralistic alignment [142] to prevent vocal minorities from dominating consensus outcomes or sidelining marginalized voices.

Beyond new ML algorithms, advancing community-centered AI also requires new interactive tools and policies that support and sustain community participation, considering its holistic nature across the development pipeline (§3). While such tools and policies may lie outside traditional ML research, AI researchers can advance these areas through collaboration with human-computer interaction (HCI) researchers and policymakers. For interactive tools, HCI researchers have deep roots in participatory design [111] and computer-supported cooperative work [1], with rich literature on how to help communities translate local norms into technical specifications [50, 177], reduce participation fatigue [25, 160], and counter adversarial manipulation or intra-community power imbalances [77, 90]. Working together, AI and HCI researchers can adapt these insights to develop interactive tools and processes for community-centered AI [84–86]. For policies, researchers and policymakers can establish institutional incentives, such as in public procurement standards or funding criteria [63], that reward authentic community-centered AI by referencing our framework (§4) to assess the depth of community engagement and deter tokenistic "community-washing" [137].

6 Counterarguments and Objections

This section addresses common skepticism about community-centered AI in a Q&A format.

Q1: *Would community-centered AI risk homogenizing what a community wants?*

A1: While community-centered AI emphasizes the needs and values of a community as a whole, it explicitly recognizes that members within a community may hold diverse views. These differences may arise from the varying roles, responsibilities, and power among members. This paper highlights the value of disagreements (§2.3) and underscores the decision-making process as a key dimension for examining how collective decisions are made (§4). Developing methods to capture pluralistic views and empower marginalized voices is an active and important area of inquiry [46, 142].

Q2: *Would community-centered AI risk conflicting with broader ethical standards?*

A2: While community-centered AI focuses on the preferences of individual communities, this does not imply unrestricted freedom that could result in harm or violate fundamental human rights [155]. Analogous to user-centered design, where prioritizing user needs does not mean ignoring all other design constraints and desiderata, community-centered AI should operate within the bounds of broader societal regulations and governance frameworks, such as the United Nations' Global Dialogue on AI Governance [154].

Q3: *Would community-centered AI be inefficient?*

A3: A similar critique is often directed at democracy, which is considered "inefficient" compared to authoritarianism. However, the checks and balances inherent in democratic processes lead to decisions that are ultimately more sustainable and legitimate [3, 45, 102]. Similarly, community-centered AI emphasizes process-oriented benefits and prioritizes long-term alignment and legitimacy over short-term cost savings. Meanwhile, research continues to explore ways to prioritize community input and attention, thereby maximizing the effectiveness of community engagement [18, 78, 84, 86].

Q4: *Would community-centered AI be difficult to scale?*

A4: In contrast to general-purpose AI, which prioritizes generality and scale, community-centered AI is inherently focused on specificity and locality. While a specific community-tailored AI system may be localized, the underlying algorithms, tools, and infrastructure required to build it (§5.2) may generalize across communities. Community-centered AI thus offers a rich research frontier for both researchers interested in concrete local impact and those pursuing generalizable contributions.

7 Related Concepts

7.1 Human-Centered AI

Community-centered AI falls within the broader framework of human-centered AI, an approach to AI that prioritizes human needs, values, and well-being throughout the development and deployment of AI systems [133, 134, 144]. James Landay, Director of Stanford HAI, describes human-centered AI as encompassing three levels: user-centered, community-centered, and society-centered [143]. While user-centered design and society-level studies of AI have attracted significant research attention (e.g., [6, 173]), community-centered AI remains relatively underexplored. This paper calls for increased research on human-centered AI at the community level.

To better distinguish community-centered AI from society-centered AI, we can draw on German sociologist Ferdinand Tönnies' classic distinction between *Gemeinschaft* and *Gesellschaft*, often translated as community and society, respectively [150]. Through this lens, society-centered AI concerns populations bound by formal state institutions and governed through legal frameworks, such as national sovereign AI initiatives [98] or international AI regulations [6]. In contrast, community-centered AI involves groups bound by shared identities, cultural norms, or affective ties (§1), such as a Māori tribe developing its own LLM and governing data through communal norms [54].

7.2 Participatory AI

Community-centered AI also falls under the umbrella of participatory AI, which draws on participatory design to actively involve stakeholders throughout the AI development process [28].[3] Within participatory AI, rather than engaging end users at the individual level [168], community-centered AI focuses on involving specific impacted communities at the collective level. This focus introduces unique challenges in participation, such as accounting for conflicts among community members and building consensus for collective decisions. Moreover, given the overlapping nature of communities, where a single person may belong to multiple communities with competing interests, community-level participation faces the added challenge of navigating both intra- and inter-community tensions.

Community-level participation also differs from society-level participation in who is represented and how decisions are made. Regarding representation, society-level participation often flattens diverse communities into a diffuse public, such as gathering "societal input" by sampling crowdworkers solely by demographic attributes [76], without examining which communities they represent. Regarding decision-making, whereas society-level efforts like democratic AI [118] focus on democratic mechanisms, community-centered AI need not be purely democratic. For example, Wikipedia policy explicitly states that Wikipedia is not a democracy; instead, decisions are made by consensus [165].

Taken together (§7.1–7.2), these distinctions show that community-centered AI cannot simply be subsumed under user-level or society-level paradigms. By foregrounding the community as a distinct unit of analysis and participation, we call on researchers, practitioners, and policymakers to advance AI systems that genuinely align with and empower the communities they impact.

[3] For how participatory design differs from other design traditions, such as value-sensitive design [40], see [28].

Acknowledgments and Disclosure of Funding

The funding for this research was provided by the National Institute of Standards and Technology (NIST) (ror.org/05xpvk416), Carnegie Mellon University (ror.org/05x2bcf33) AI Measurement Science and Engineering Center, and the Block Center for Technology and Society. Tzu-Sheng Kuo was also supported by the K&L Gates Presidential Fellowship in Ethics and Computational Technologies. We thank Jason Hong, Aaron Halfaker, Hong Shen, Michael Bernstein, Aniket Kittur, and Tiffany Chih for their early feedback on the outline of this paper and the development of the community-centered AI framework.

References

[1] Mark S Ackerman. 2000. The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. *Human–Computer Interaction* 15, 2-3 (2000), 179–203.

[2] Ali Alkhatib and Michael Bernstein. 2019. Street-Level Algorithms: A Theory at the Gaps Between Policy and Decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300760

[3] David Altman, Patrick Flavin, and Benjamin Radcliff. 2017. Democratic Institutions and Subjective Well-being. *Political Studies* 65, 3 (2017), 685–704.

[4] Sherry R. Arnstein. 1969. A Ladder Of Citizen Participation. *Journal of the American Institute of Planners* 35, 4 (1969), 216–224. doi:10.1080/01944366908977225

[5] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073* (2022).

[6] Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Danielle Goldfarb, et al. 2025. International AI Safety Report. *arXiv preprint arXiv:2501.17805* (2025).

[7] Stevie Bergman, Nahema Marchal, John Mellor, Shakir Mohamed, Iason Gabriel, and William Isaac. 2024. STELA: A Community-Centred Approach to Norm Elicitation for AI Alignment. *Scientific Reports* 14, 1 (2024), 6616.

[8] Eshta Bhardwaj, Harshit Gujral, Siyi Wu, Ciara Zogheib, Tegan Maharaj, and Christoph Becker. 2024. Machine learning data practices through a data curation lens: An evaluation framework. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1055–1067. doi:10.1145/3630106.3658955

[9] Bhavana Bheem, Samuel Gerstein, Karrie Karahalios, and Madelyn Rose Sanfilippo. 2026. AI Localism: Counter Imaginaries from City AI Governance. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 9. 435–446.

[10] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. 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* (Arlington, VA, USA) *(EAAMO '22)*. Association for Computing Machinery, New York, NY, USA, Article 6, 8 pages. doi:10.1145/3551624.3555290

[11] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. 173–184.

[12] Henry Campbell Black, Bryan A Garner, Becky R McDaniel, David W Schultz, and West Publishing Company. 1999. *Black's Law Dictionary*. Vol. 196. West Group St. Paul, MN.

[13] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman,

Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the Opportunities and Risks of Foundation Models. *ArXiv* (2021). `https://crfm.stanford.edu/assets/report.pdf`

[14] John G Bruhn. 2005. *The Sociology of Community Connections*. Springer.

[15] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91. `https://proceedings.mlr.press/v81/buolamwini18a.html`

[16] Rochelle Campbell-Esen, Jamie Mahoney, and Reem Talhouk. 2026. Co-Designing Islamic AI Ethics: Insights from the UK Muslim Community. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, Article 930, 18 pages. doi:`10.1145/3772318.3790413`

[17] Brian J Chen and Jacob Metcalf. 2024. Explainer: A Sociotechnical Approach to AI Policy. *Data & Society* (2024).

[18] Quan Ze Chen and Amy X Zhang. 2023. Judgment sieve: Reducing uncertainty in group judgments through interventions targeting ambiguity versus disagreement. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–26.

[19] Hao-Fei Cheng, Logan Stapleton, Anna Kawakami, Venkatesh Sivaraman, Yanghuidi Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 162, 22 pages. doi:`10.1145/3491102.3501831`

[20] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (2017), 153–163. doi:`10.1089/big.2016.0047`

[21] Patricia Hill Collins. 2022. *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*. Routledge.

[22] Ned Cooper, Tiffanie Horne, Gillian R Hayes, Courtney Heldreth, Michal Lahav, Jess Holbrook, and Lauren Wilcox. 2022. A Systematic Review and Thematic Analysis of Community-Collaborative Approaches to Computing Research. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 73, 18 pages. doi:`10.1145/3491102.3517716`

[23] Eric Corbett, Remi Denton, and Sheena Erete. 2023. Power and Public Participation in AI. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Boston, MA, USA) *(EAAMO '23)*. Association for Computing Machinery, New York, NY, USA, Article 8, 13 pages. doi:`10.1145/3617694.3623228`

[24] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS, Canada) *(KDD '17)*. Association for Computing Machinery, New York, NY, USA, 797–806. doi:`10.1145/3097983.3098095`

[25] Dan Cosley, Dan Frankowski, Loren Terveen, and John Riedl. 2007. SuggestBot: using intelligent task routing to help people find work in Wikipedia. In *Proceedings of the 12th international conference on Intelligent user interfaces*. 32–41.

[26] Humphrey Curtis, Adam D G Jenkins, Alistair Gentry, Sioban Zacharek, Sally McVicker, Timothy Neate, and Filip Bircanin. 2026. "Computer Says No": Disabled Welfare Experiences and Envisioned Futures Under AI Governance. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, Article 29, 25 pages. doi:10.1145/3772318.3791025

[27] Annelien De Dijn. 2020. *Freedom: An Unruly History*. Harvard University Press.

[28] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. 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* (Boston, MA, USA) *(EAAMO '23)*. Association for Computing Machinery, New York, NY, USA, Article 37, 23 pages. doi:10.1145/3617694.3623261

[29] Wesley Hanwen Deng, Wang Claire, Howard Ziyu Han, Jason I. Hong, Kenneth Holstein, and Motahhare Eslami. 2025. WeAudit: Scaffolding User Auditors and AI Practitioners in Auditing Generative AI. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW521 (Oct. 2025), 35 pages. doi:10.1145/3757702

[30] Remi Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. 2020. Bringing the people back in: Contesting benchmark machine learning datasets. *arXiv preprint arXiv:2007.07399* (2020).

[31] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Model Alignment as Prospect Theoretic Optimization. In *Forty-first International Conference on Machine Learning*. https://openreview.net/forum?id=iUwHnoENnl

[32] Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. Macmillan+ ORM.

[33] Jenny Fan and Amy X. Zhang. 2020. Digital Juries: A Civics-Oriented Approach to Platform Governance. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376293

[34] Sina Fazelpour and Will Fleisher. 2025. The Value of Disagreement in AI Design, Evaluation, and Alignment. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 2138–2150. doi:10.1145/3715275.3732146

[35] K. J. Kevin Feng, Tzu-Sheng Kuo, Quan Ze (Jim) Chen, Inyoung Cheong, Kenneth Holstein, and Amy X. Zhang. 2026. PolicyPad: Collaborative Prototyping of LLM Policies. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, Article 39, 25 pages. doi:10.1145/3772318.3791689

[36] K. J. Kevin Feng, Rock Yuren Pang, Tzu-Sheng Kuo, Amy Winecoff, Emily Tseng, David Gray Widder, Harini Suresh, Katharina Reinecke, and Amy X. Zhang. 2025. Sociotechnical AI Governance: Challenges and Opportunities for HCI. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 800, 6 pages. doi:10.1145/3706599.3706747

[37] Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024. Modular Pluralism: Pluralistic Alignment via Multi-LLM Collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 4151–4171. doi:10.18653/v1/2024.emnlp-main.240

[38] Casey Fiesler, Jialun Jiang, Joshua McCann, Kyle Frye, and Jed Brubaker. 2018. Reddit Rules! Characterizing an Ecosystem of Governance. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.

[39] James S Fishkin. 1991. *Democracy and Deliberation: New Directions for Democratic Reform*. Yale University Press.

[40] Batya Friedman. 1996. Value-sensitive design. *interactions* 3, 6 (1996), 16–23.

[41] Archon Fung. 2006. Varieties of Participation in Complex Governance. *Public Administration Review* 66, s1 (2006), 66–75. doi:10.1111/j.1540-6210.2006.00667.x

[42] Seeta Peña Gangadharan. 2021. Digital exclusion: A politics of refusal. *Digital technology and democratic theory* 4 (2021), 113–40.

[43] Lan Gao, Abani Ahmed, Oscar Chen, Margaux Reyl, Zayna Cheema, Nick Feamster, Chenhao Tan, Kurt Thomas, and Marshini Chetty. 2026. Governance of AI-Generated Content: A Case Study on Social Media Platforms. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, Article 1604, 22 pages. doi:10.1145/3772318.3790415

[44] Zihan Gao, Yifei Xu, and Jacob Thebault-Spieker. 2026. Localbench: Benchmarking LLMs on county-level local knowledge and reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 38487–38495.

[45] John Gerring, Carl Henrik Knutsen, and Jonas Berge. 2022. Does Democracy Matter? *Annual Review of Political Science* 25 (2022), 357–375.

[46] Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 115, 19 pages. doi:10.1145/3491102.3502004

[47] Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. 2021. The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 388, 14 pages. doi:10.1145/3411764.3445423

[48] Luke Guerdan, Devansh Saxena, Stevie Chancellor, Zhiwei Steven Wu, and Kenneth Holstein. 2025. Measurement as Bricolage: Examining How Data Scientists Construct Target Variables for Predictive Modeling Tasks. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW447 (Oct. 2025), 37 pages. doi:10.1145/3757628

[49] Jürgen Habermas. 1975. *Legitimation Crisis*. Vol. 519. Beacon Press.

[50] Aaron Halfaker and R. Stuart Geiger. 2020. ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 148 (Oct. 2020), 37 pages. doi:10.1145/3415219

[51] Aaron Halfaker, Aniket Kittur, and John Riedl. 2011. Don't bite the newbies: how reverts affect the quantity and quality of Wikipedia work. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration* (Mountain View, California) *(WikiSym '11)*. Association for Computing Machinery, New York, NY, USA, 163–172. doi:10.1145/2038558.2038585

[52] Aaron L Halfaker, Tzu-Sheng Kuo, Ciell Brusse, Kenneth Holstein, and Haiyi Zhu. 2025. Collective Meaning Cascades but Strange Ducks Swim Upstream: Facilitating Collective Meaning-making through Co-development of AI Models. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 660, 8 pages. doi:10.1145/3706599.3706683

[53] Hamna Hamna, Gayatri Bhat, Sourabrata Mukherjee, Faisal M. Lalani, Evan Hadfield, Divya Siddarth, Kalika Bali, and Sunayana Sitaram. 2026. Building Benchmarks from the Ground Up: Community-Centered Evaluation of LLMs in Healthcare Chatbot Settings. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, Article 23, 19 pages. doi:10.1145/3772318.3791172

[54] Karen Hao. 2025. *Empire of AI: Dreams and Nightmares in Sam Altman's OpenAI*. Penguin Press, New York, NY.

[55] MD Romael Haque, Devansh Saxena, Katy Weathington, Joseph Chudzik, and Shion Guha. 2024. Are We Asking the Right Questions?: Designing for Community Stakeholders' Interactions with AI in Policing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 301, 20 pages. doi:10.1145/3613904.3642738

[56] Donna Haraway. 2013. Situated knowledges: The science question in feminism and the privilege of partial perspective. In *Women, science, and technology*. Routledge, 455–472.

[57] Sandra G Harding. 2004. *The feminist standpoint theory reader: Intellectual and political controversies*. Psychology Press.

[58] Gillian R Hayes. 2014. Knowing by doing: action research as an approach to HCI. In *Ways of Knowing in HCI*. Springer, 49–68.

[59] Wayne Holmes, Kaska Porayska-Pomsta, Ken Holstein, Emma Sutherland, Toby Baker, Simon Buckingham Shum, Olga C Santos, Mercedes T Rodrigo, Mutlu Cukurova, Ig Ibert Bittencourt, et al. 2022. Ethics of AI in Education: Towards a Community-Wide Framework. *International Journal of Artificial Intelligence in Education* 32, 3 (2022), 504–526.

[60] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=nZeVKeeFYf9

[61] Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I. Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. 2024. Collective Constitutional AI: Aligning a Language Model with Public Input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1395–1417. doi:10.1145/3630106.3658979

[62] Robert Huckfeldt, Paul E Johnson, and John Sprague. 2004. *Political Disagreement: The Survival of Diverse Opinions within Communication Networks*. Cambridge University Press.

[63] Anna Ida Hudig, Emma Kallina, and Jatinder Singh. 2026. "It's Just a Wild, Wild West": Harnessing Public Procurement as an AI Governance Mechanism. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, Article 35, 22 pages. doi:10.1145/3772318.3791968

[64] Julie Hui, Kristin Seefeldt, Lutalo Sanifu, Christie Baer, Jeanette Szomstein, and Tawanna R Dillahunt. 2024. "I was able to give her the confidence": Reciprocal Capacity Building in a Community-based Program for Digital Engagement. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 927, 13 pages. doi:10.1145/3613904.3642209

[65] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Trans. Comput.-Hum. Interact.* 26, 5, Article 31 (July 2019), 35 pages. doi:10.1145/3338243

[66] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. 2023. AI alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852* (2023).

[67] Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 306–316. doi:10.1145/3351095.3372829

[68] Cathryn Johnson, Timothy J Dowd, and Cecilia L Ridgeway. 2006. Legitimacy as a Social Process. *Annu. Rev. Sociol.* 32, 1 (2006), 53–78.

[69] Nari Johnson, Sanika Moharana, Christina Harrington, Nazanin Andalibi, Hoda Heidari, and Motahhare Eslami. 2024. The Fall of an Algorithm: Characterizing the Dynamics Toward Abandonment. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 337–358. doi:10.1145/3630106.3658910

[70] Nari Johnson, Deepthi Sudharsan, Hamna, Samantha Dalal, Theo Holroyd, Anja Thieme, Hoda Heidari, Daniela Massiceti, Jennifer Wortman Vaughan, and Cecily Morrison. 2026. Evaluating AI-Generated Images of Cultural Artifacts with Community-Informed Rubrics. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency*. 714–774.

[71] Nithish Kannen, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji B Dieng, Pushpak Bhattacharyya, and Shachi Dave. 2024. Beyond aesthetics: Cultural competence in text-to-image models. *Advances in Neural Information Processing Systems* 37 (2024), 13716–13747.

[72] Anna Kawakami, Amanda Coston, Haiyi Zhu, Hoda Heidari, and Kenneth Holstein. 2024. The Situate AI Guidebook: Co-Designing a Toolkit to Support Multi-Stakeholder, Early-stage Deliberations Around Public Sector AI Proposals. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 749, 22 pages. doi:10.1145/3613904.3642849

[73] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 52, 18 pages. doi:10.1145/3491102.3517439

[74] Anna Kawakami, Jordan Taylor, Sarah Fox, Haiyi Zhu, and Kenneth Holstein. 2026. AI Failure Loops in Devalued Work: The confluence of overconfidence in AI and underconfidence in worker expertise. *Big Data & Society* 13, 1 (2026), 20539517261424164.

[75] Anna Kawakami, Haiyi Zhu, and Kenneth Holstein. 2026. AI Measurement as a Collaborative Design Practice: A Multidisciplinary Synthesis and Future Directions. *Available at SSRN 6276543* (2026).

[76] Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Michael Bean, Katerina Margatina, Rafael Mosquera, Juan Manuel Ciro, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=DFr5hteojx

[77] Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. 2007. He says, she says: conflict and coordination in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 453–462. doi:10.1145/1240624.1240698

[78] Vinay Koshy, Frederick Choi, Yi-Shyuan Chiang, Hari Sundaram, Eshwar Chandrasekharan, and Karrie Karahalios. 2025. Venire: A machine learning-guided panel review system for community content moderation. *Proceedings of the ACM on Human-Computer Interaction* 9, 7 (2025), 1–35.

[79] Yasmine Kotturi, Julie Hui, TJ Johnson, Lutalo Sanifu, and Tawanna R. Dillahunt. 2024. Sustaining Community-Based Research in Computing: Lessons from Two Tech Capacity Building Initiatives for Local Businesses. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 178 (April 2024), 31 pages. doi:10.1145/3641017

[80] P. M. Krafft, Meg Young, Michael Katell, Jennifer E. Lee, Shankar Narayan, Micah Epstein, Dharma Dailey, Bernease Herman, Aaron Tam, Vivian Guetler, Corinne Bintz, Daniella Raz, Pa Ousman Jobe, Franziska Putz, Brian Robick, and Bissan Barghouti. 2021. An Action-Oriented AI Policy Toolkit for Technology Audits by Community Advocates and Activists. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 772–781. doi:10.1145/3442188.3445938

[81] Robert E Kraut and Paul Resnick. 2012. *Building Successful Online Communities: Evidence-Based Social Design*. Mit Press.

[82] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, F. Pereira, C.J. Burges, L. Bottou, and K. Weinberger (Eds.), Vol. 25. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[83] Blaine Kuehnert, Rachel Kim, Jodi Forlizzi, and Hoda Heidari. 2025. The "Who", "What", and "How" of Responsible AI Governance: A Systematic Review and Meta-Analysis of (Actor, Stage)-Specific Tools. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 2991–3005. doi:10.1145/3715275.3732191

[84] Tzu-Sheng Kuo, Quan Ze Chen, Amy X. Zhang, Jane Hsieh, Haiyi Zhu, and Kenneth Holstein. 2025. PolicyCraft: Supporting Collaborative and Participatory Policy Design through Case-Grounded Deliberation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 805, 24 pages. doi:10.1145/3706598.3713865

[85] Tzu-Sheng Kuo, Aaron Lee Halfaker, Zirui Cheng, Jiwoo Kim, Meng-Hsin Wu, Tongshuang Wu, Kenneth Holstein, and Haiyi Zhu. 2024. Wikibench: Community-Driven Data Curation for AI Evaluation on Wikipedia. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 193, 24 pages. doi:10.1145/3613904.3642278

[86] Tzu-Sheng Kuo, Sophia Liu, Quan Ze Chen, Joseph Seering, Amy X. Zhang, Haiyi Zhu, and Kenneth Holstein. 2026. Botender: Supporting Communities in Collaboratively Designing AI Agents through Case-Based Provocations. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, Article 274, 31 pages. doi:10.1145/3772318.3790500

[87] Tzu-Sheng Kuo, Hong Shen, Jisoo Geum, Nev Jones, Jason I. Hong, Haiyi Zhu, and Kenneth Holstein. 2023. Understanding Frontline Workers' and Unhoused Individuals' Perspectives on AI Used in Homeless Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 860, 17 pages. doi:10.1145/3544548.3580882

[88] Lin Kyi, Amruta Mahuli, M. Six Silberman, Reuben Binns, Jun Zhao, and Asia J. Biega. 2025. Governance of Generative AI in Creative Work: Consent, Credit, Compensation, and Beyond. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 197, 16 pages. doi:10.1145/3706598.3713799

[89] Michelle S. Lam, Mitchell L. Gordon, Danaë Metaxa, Jeffrey T. Hancock, James A. Landay, and Michael S. Bernstein. 2022. End-User Audits: A System Empowering Communities to Lead Large-Scale Investigations of Harmful Algorithmic Behavior. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 512 (Nov. 2022), 34 pages. doi:10.1145/3555625

[90] Cliff Lampe and Paul Resnick. 2004. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 543–550.

[91] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324. doi:10.1109/5.726791

[92] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D. Procaccia. 2019. WeBuildAI: Participatory Framework for Algorithmic Governance. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 181 (Nov. 2019), 35 pages. doi:10.1145/3359283

[93] Leon Leibmann, Galen Weld, Amy X Zhang, and Tim Althoff. 2025. Reddit Rules and Rulers: Quantifying the Link Between Rules and Perceptions of Governance across Thousands of Communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 19. 1098–1121.

[94] Chu Li, Katrina Oi Yau Ma, Michael Saugstad, Kie Fujii, Molly Delaney, Yochai Eisenberg, Delphine Labbé, Judy L Shanley, Devon Snyder, Florian P P Thomas, and Jon E. Froehlich. 2024. "I never realized sidewalks were a big deal": A Case Study of a Community-Driven Sidewalk Accessibility Assessment using Project Sidewalk. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 969, 18 pages. doi:10.1145/3613904.3642003

[95] Xingyu Li, Alexandra Teixeira Riggs, Zhiming Dai, Crystal Byrd Farmer, Kalia G Morrison, and Noura Howell. 2026. Informal Embodied Auditing: Exploring Facial Emotion AI (FEAI) through Community Workshops. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, Article 166, 21 pages. doi:10.1145/3772318.3790863

[96] Selma C Liberato, Julie Brimblecombe, Jan Ritchie, Megan Ferguson, and John Coveney. 2011. Measuring Capacity Building in Communities: A Review of the Literature. *BMC public health* 11, 1 (2011), 850.

[97] Hongjin Lin, Naveena Karusala, Chinasa T Okolo, Catherine D'Ignazio, and Krzysztof Z Gajos. 2024. "Come to us first": Centering Community Organizations in Artificial Intelligence for Social Good Partnerships. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–28.

[98] Yen-Ting Lin and Yun-Nung Chen. 2023. Taiwan LLM: Bridging the linguistic divide with a culturally aligned language model. *arXiv preprint arXiv:2311.17487* (2023).

[99] Seymour Martin Lipset. 1959. Some Social Requisites of Democracy: Economic Development and Political Legitimacy. *American political science review* 53, 1 (1959), 69–105.

[100] Travis Lloyd, Jennah Gosciak, Tung Nguyen, and Mor Naaman. 2025. AI Rules? Characterizing Reddit Community Policies Towards AI-Generated Content. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 9, 19 pages. doi:10.1145/3706598.3713292

[101] Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3313831.3376727

[102] Reinet Loubser and Cindy Steenekamp. 2017. Democracy, Well-being, and Happiness: A 10-nation Study. *Journal of Public Affairs* 17, 1-2 (2017), e1646.

[103] Yao Lyu, Jie Cai, and John M. Carroll. 2025. A Systematic Literature Review of Infrastructure Studies in SIGCHI. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW364 (Oct. 2025), 41 pages. doi:10.1145/3757545

[104] Richard Martinez and Kurt Squire. 2024. Engaging recently incarcerated and gang affiliated Black and Latino/a young adults in designing social collocated applications for mixed reality smart glasses through community-based participatory design workshops.. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 871, 17 pages. doi:10.1145/3613904.3642895

[105] Danaë Metaxa, Joon Sung Park, Ronald E. Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, and Christian Sandvig. 2021. Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. *Foundations and Trends in Human-Computer Interaction* 14, 4 (11 2021), 272–344. doi:10.1561/1100000083

[106] Joan G Miller and David M Bersoff. 1992. Culture and moral judgment: How are conflicts between justice and interpersonal responsibilities resolved? *Journal of personality and social psychology* 62, 4 (1992), 541.

[107] Swati Mishra and Jeffrey M Rzeszotarski. 2021. Designing Interactive Transfer Learning Tools for ML Non-Experts. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 364, 15 pages. doi:10.1145/3411764.3445096

[108] Erina Seh-Young Moon and Shion Guha. 2024. A Human-Centered Review of Algorithms in Homelessness Research. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 870, 15 pages. doi:10.1145/3613904.3642392

[109] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q. Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3290605.3300356

[110] Michael Muller, Christine T. Wolf, Josh Andres, Michael Desmond, Narendra Nath Joshi, Zahra Ashktorab, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Evelyn Duesterwald, and Casey Dugan. 2021. Designing Ground Truth and the Social Life of Labels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 94, 16 pages. doi:10.1145/3411764.3445402

[111] Michael J Muller and Sarah Kuhn. 1993. Participatory design. *Commun. ACM* 36, 6 (1993), 24–28.

[112] Rashid Mushkani, Shravan Nayak, Hugo Berard, Allison Cohen, Shin Koseki, and Hadrien Bertrand. 2025. LIVS: A Pluralistic Alignment Dataset for Inclusive Public Spaces. In *Forty-second International Conference on Machine Learning*. https://openreview.net/forum?id=Spoe53kbj9

[113] Luis Oala, Manil Maskey, Lilith Bat-Leah, Alicia Parrish, Nezihe Merve Gürel, Tzu-Sheng Kuo, Yang Liu, Rotem Dror, Danilo Brajovic, Xiaozhe Yao, Max Bartolo, William A Gaviria Rojas, Ryan Hileman, Rainier Aliment, Michael W. Mahoney, Meg Risdal, Matthew Lease, Wojciech Samek, Debojyoti Dutta, Curtis G Northcutt, Cody Coleman, Braden Hancock, Bernard Koch, Girmaw Abebe Tadesse, Bojan Karlaš, Ahmed Alaa, Adji Bousso Dieng, Natasha Noy, Vijay Janapa Reddi, James Zou, Praveen Paritosh,

Mihaela van der Schaar, Kurt Bollacker, Lora Aroyo, Ce Zhang, Joaquin Vanschoren, Isabelle Guyon, and Peter Mattson. 2024. DMLR: Data-centric Machine Learning Research - Past, Present and Future. *Journal of Data-centric Machine Learning Research* (2024). https://openreview.net/forum?id=2kpu78QdeE Featured Certification, Survey Certification.

[114] Executive Office of the President. 2023. Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. *Federal Register* 88 (2023), 75191–75226.

[115] Chinasa T. Okolo, Srujana Kamath, Nicola Dell, and Aditya Vashistha. 2021. "It cannot do all of my work": Community Health Worker Perceptions of AI-Enabled Mobile Health Applications in Rural India. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 701, 20 pages. doi:10.1145/3411764.3445420

[116] Elinor Ostrom. 1990. Governing the Commons: The Evolution of Institutions for Collective Action. *Cambridge University Press* (1990).

[117] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 27730–27744. https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf

[118] Aviv Ovadya, Kyle Redman, Luke Thorburn, Quan Ze Chen, Oliver Smith, Flynn Devine, Andrew Konya, Smitha Milli, Manon Revel, Kevin Feng, Amy X Zhang, Bilva Chandra, Michiel A. Bakker, and Atoosa Kasirzadeh. 2025. Position: Democratic AI is Possible. The Democracy Levels Framework Shows How It Might Work.. In *Proceedings of the 42nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 267)*, Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (Eds.). PMLR, 81930–81961. https://proceedings.mlr.press/v267/ovadya25a.html

[119] Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 39–48. doi:10.1145/3287560.3287567

[120] Organizers Of Queerinai, Anaelia Ovalle, Arjun Subramonian, Ashwin Singh, Claas Voelcker, Danica J. Sutherland, Davide Locatelli, Eva Breznik, Filip Klubicka, Hang Yuan, Hetvi J, Huan Zhang, Jaidev Shriram, Kruno Lehman, Luca Soldaini, Maarten Sap, Marc Peter Deisenroth, Maria Leonor Pacheco, Maria Ryskina, Martin Mundt, Milind Agarwal, Nyx Mclean, Pan Xu, A Pranav, Raj Korpan, Ruchira Ray, Sarah Mathew, Sarthak Arora, St John, Tanvi Anand, Vishakha Agrawal, William Agnew, Yanan Long, Zijie J. Wang, Zeerak Talat, Avijit Ghosh, Nathaniel Dennler, Michael Noseworthy, Sharvani Jha, Emi Baylor, Aditya Joshi, Natalia Y. Bilenko, Andrew Mcnamara, Raphael Gontijo-Lopes, Alex Markham, Evyn Dong, Jackie Kay, Manu Saraswat, Nikhil Vytla, and Luke Stark. 2023. Queer In AI: A Case Study in Community-Led Participatory AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1882–1895. doi:10.1145/3593013.3594134

[121] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[122] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=HPuSIXJaa9

[123] Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the Everything in the Whole Wide World Benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, J. Vanschoren and S. Yeung (Eds.), Vol. 1. https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/084b6fbb10729ed4da8c3d3f5a3ae7c9-Paper-round2.pdf

[124] Anka Reuel, Benjamin Bucknall, Stephen Casper, Timothy Fist, Lisa Soder, Onni Aarne, Lewis Hammond, Lujain Ibrahim, Alan Chan, Peter Wills, Markus Anderljung, Ben Garfinkel, Lennart Heim, Andrew Trask, Gabriel Mukobi, Rylan Schaeffer, Mauricio Baker, Sara Hooker, Irene Solaiman, Sasha Luccioni, Nitarshan Rajkumar, Nicolas Moës, Jeffrey Ladish, David Bau, Paul Bricman, Neel Guha, Jessica

Newman, Yoshua Bengio, Tobin South, Alex Pentland, Sanmi Koyejo, Mykel Kochenderfer, and Robert Trager. 2025. Open Problems in Technical AI Governance. *Transactions on Machine Learning Research* (2025). https://openreview.net/forum?id=1nO4qFMiS0 Survey Certification.

[125] Anka Reuel, Lisa Soder, Benjamin Bucknall, and Trond Arne Undheim. 2024. Position: Technical Research and Talent is Needed for Effective AI Governance. In *Forty-first International Conference on Machine Learning*. https://openreview.net/forum?id=Be2B6f0ps1

[126] Samantha Robertson, Tonya Nguyen, and Niloufar Salehi. 2021. Modeling Assumptions Clash with the Real World: Transparency, Equity, and Community Challenges for Student Assignment Algorithms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 589, 14 pages. doi:10.1145/3411764.3445748

[127] Philippe Rochat, Maria DG Dias, Guo Liping, Tanya Broesch, Claudia Passos-Ferreira, Ashley Winning, and Britt Berg. 2009. Fairness in Distributive Justice by 3-and 5-year-olds Across Seven Cultures. *Journal of Cross-Cultural Psychology* 40, 3 (2009), 416–442.

[128] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 1668–1678. doi:10.18653/v1/P19-1163

[129] Devansh Saxena, Karla Badillo-Urquiola, Pamela J. Wisniewski, and Shion Guha. 2020. A Human-Centered Review of Algorithms used within the U.S. Child Welfare System. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3313831.3376229

[130] Stefan Schaal. 1996. Learning from demonstration. *Advances in neural information processing systems* 9 (1996).

[131] John Scott and Gordon Marshall. 2009. *A Dictionary of Sociology*. Oxford University Press.

[132] Joseph Seering, Juan Pablo Flores, Saiph Savage, and Jessica Hammer. 2018. The Social Roles of Bots: Evaluating Impact of Bots on Discussions in Online Communities. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 157 (Nov. 2018), 29 pages. doi:10.1145/3274426

[133] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504.

[134] Ben Shneiderman. 2021. Human-centered AI. *Issues in Science and Technology* 37, 2 (2021), 56–61.

[135] Saad Siddiqui, Lujain Ibrahim, Kristy Loke, Stephen Clare, Marianne Lu, Aris Richardson, Conor McGlynn, and Jeffrey Ding. 2025. Promising Topics for US–China Dialogues on AI Risks and Governance. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 1211–1223. doi:10.1145/3715275.3732080

[136] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[137] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2022. Participation is not a design fix for machine learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–6.

[138] C. Estelle Smith, Bowen Yu, Anjali Srivastava, Aaron Halfaker, Loren Terveen, and Haiyi Zhu. 2020. 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* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376783

[139] Jaemarie Solyst, Cindy Peng, Wesley Hanwen Deng, Praneetha Pratapa, Amy Ogan, Jessica Hammer, Jason Hong, and Motahhare Eslami. 2025. Investigating Youth AI Auditing. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 2098–2111. doi:10.1145/3715275.3732142

[140] Jaemarie Solyst, Shixian Xie, Ellia Yang, Angela E.B. Stewart, Motahhare Eslami, Jessica Hammer, and Amy Ogan. 2023. "I Would Like to Design": Black Girls Analyzing and Ideating Fair and Accountable AI. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 452, 14 pages. doi:10.1145/3544548.3581378

[141] Jaemarie Solyst, Ellia Yang, Shixian Xie, Jessica Hammer, Amy Ogan, and Motahhare Eslami. 2024. Children's Overtrust and Shifting Perspectives of Generative AI. *arXiv preprint arXiv:2404.14511* (2024).

[142] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: A Roadmap to Pluralistic Alignment. In *Forty-first International Conference on Machine Learning*. https://openreview.net/forum?id=gQpBnRHwxM

[143] Stanford Institute for Human-Centered Artificial Intelligence. 2026. AI for Good Isn't Good Enough: A Call for Human-Centered AI. https://hai.stanford.edu/events/ai-good-isnt-good-enough-call-human-centered-ai Accessed: 2026-05-04.

[144] Stanford Institute for Human-Centered Artificial Intelligence. 2026. What is Human-Centered AI? https://hai.stanford.edu/ai-definitions/what-is-human-centered-ai Accessed: 2026-05-04.

[145] Logan Stapleton, Min Hun Lee, Diana Qing, Marya Wright, Alexandra Chouldechova, Ken Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. Imagining new futures beyond predictive systems in child welfare: A qualitative study with impacted stakeholders. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1162–1177. doi:10.1145/3531146.3533177

[146] Roy Suddaby, Alex Bitektine, and Patrick Haack. 2017. Legitimacy. *Academy of Management Annals* 11, 1 (2017), 451–478.

[147] Harini Suresh, Emily Tseng, Meg Young, Mary Gray, Emma Pierson, and Karen Levy. 2024. Participation in the age of foundation models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1609–1621. doi:10.1145/3630106.3658992

[148] Ningjing Tang, Jiayin Zhi, Tzu-Sheng Kuo, Calla Kainaroi, Jeremy J. Northup, Kenneth Holstein, Haiyi Zhu, Hoda Heidari, and Hong Shen. 2024. AI Failure Cards: Understanding and Supporting Grassroots Efforts to Mitigate AI Failures in Homeless Services. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 713–732. doi:10.1145/3630106.3658935

[149] Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. 2024. AI can help humans find common ground in democratic deliberation. *Science* 386, 6719 (2024), eadq2852. doi:10.1126/science.adq2852

[150] Ferdinand Tönnies and Charles P Loomis. 2017. *Community and Society*. Routledge.

[151] Tom R Tyler. 2003. Procedural Justice, Legitimacy, and the Effective Rule of Law. *Crime and justice* 30 (2003), 283–357.

[152] Kim M Unertl, Chris L Schaefbauer, Terrance R Campbell, Charles Senteio, Katie A Siek, Suzanne Bakken, and Tiffany C Veinot. 2016. Integrating Community-Based Participatory Research and Informatics Approaches to Improve the Engagement and Health of Underserved Populations. *Journal of the American Medical Informatics Association* 23, 1 (2016), 60–73.

[153] United Nations. 2026. *Capacity-Building*. United Nations Academic Impact. https://www.un.org/en/academic-impact/capacity-building Accessed: 2026-05-02.

[154] United Nations. 2026. Global Dialogue on AI Governance. https://www.un.org/global-dialogue-ai-governance/en Accessed: 2026-05-04.

[155] United Nations. 2026. *How AI is fuelling human trafficking*. Office on Drugs and Crime. https://www.unodc.org/unodc/en/frontpage/2026/September/how-ai-is-fuelling-human-trafficking.html Accessed: 2026-10-20.

[156] Maarten Van Mechelen, Rachel Charlotte Smith, Marie-Monique Schaper, Mariana Tamashiro, Karl-Emil Bilstrup, Mille Lunding, Marianne Graves Petersen, and Ole Sejer Iversen. 2023. Emerging Technologies in K–12 Education: A Future HCI Research Agenda. *ACM Trans. Comput.-Hum. Interact.* 30, 3, Article 47 (June 2023), 40 pages. doi:10.1145/3569897

[157] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[158] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3173574.3174014

[159] Michael Veale and Frederik Zuiderveen Borgesius. 2021. Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International* 22, 4 (2021), 97–112.

[160] Luis Von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 319–326.

[161] Hanna Wallach, Meera Desai, A. Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Nicholas J Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. 2025. Position: Evaluating Generative AI Systems Is a Social Science Measurement Challenge. In *Forty-second International Conference on Machine Learning Position Paper Track*. https://openreview.net/forum?id=1ZC4RNjqzU

[162] Nina B Wallerstein and Bonnie Duran. 2006. Using Community-Based Participatory Research to Address Health Disparities. *Health promotion practice* 7, 3 (2006), 312–323.

[163] Angelina Wang, Sayash Kapoor, Solon Barocas, and Arvind Narayanan. 2024. Against Predictive Optimization: On the Legitimacy of Decision-making Algorithms That Optimize Predictive Accuracy. *ACM J. Responsib. Comput.* 1, 1, Article 9 (March 2024), 45 pages. doi:10.1145/3636509

[164] E. Glen Weyl, Audrey Tang, and the Plurality Community. 2023. *Plurality: The Future of Collaborative Technology and Democracy*. https://github.com/pluralitybook/plurality/blob/main/contents/english

[165] Wikipedians. 2026. Wikipedia: What Wikipedia is not — Wikipedia is not a democracy. https://en.wikipedia.org/wiki/Wikipedia:What_Wikipedia_is_not#Wikipedia_is_not_a_democracy Accessed: 2026-05-04.

[166] Amy Winecoff and Miranda Bogen. 2025. Improving Governance Outcomes Through AI Documentation: Bridging Theory and Practice. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1194, 18 pages. doi:10.1145/3706598.3713814

[167] Yiwei Wu, Leah Ajmani, Nathan TeBlunthuis, and Hanlin Li. 2026. AI Didn't Start the Fire: Examining the Stack Exchange Moderator and Contributor Strike. *Proceedings of the ACM on Human-Computer Interaction* 10, 2 (2026), 1–27.

[168] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376301

[169] Kyra Yee, Uthaipon Tantipongpipat, and Shubhanshu Mishra. 2021. Image Cropping on Twitter: Fairness Metrics, their Limitations, and the Importance of Representation, Design, and Agency. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 450 (Oct. 2021), 24 pages. doi:10.1145/3479594

[170] Nur Yildirim, Changhoon Oh, Deniz Sayar, Kayla Brand, Supritha Challa, Violet Turri, Nina Crosby Walton, Anna Elise Wong, Jodi Forlizzi, James McCann, and John Zimmerman. 2023. Creating Design Resources to Scaffold the Ideation of AI Concepts. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 2326–2346. doi:10.1145/3563657.3596058

[171] Nur Yildirim, Mahima Pushkarna, Nitesh Goyal, Martin Wattenberg, and Fernanda Viégas. 2023. Investigating How Practitioners Use Human-AI Guidelines: A Case Study on the People + AI Guidebook. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 356, 13 pages. doi:10.1145/3544548.3580900

[172] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks?. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2* (Montreal, Canada) *(NIPS'14)*. MIT Press, Cambridge, MA, USA, 3320–3328.

[173] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 437, 21 pages. doi:10.1145/3544548.3581388

[174] Zhilin Zhang, Jun Zhao, Ge Wang, Sruthi Viswanathan, Tala Jo Ross, Samantha-Kaye Johnston, Diyi Liu, Hayoun Noh, Max Van Kleek, and Nigel Shadbolt. 2026. Attitudes, Imagined Roles, and Governance Boundaries for AI in Decentralized Social Media. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, Article 330, 17 pages. doi:10.1145/3772318.3790295

[175] Dora Zhao, Hannah Cha, Michael J Ryan, Angelina Wang, Rachel Baker-Ramos, Evyn-Bree Helekahi-Kaiwi, Rebecca Diego, Josiah Hester, and Diyi Yang. 2026. Whose Knowledge Counts? Co-Designing Community-Centered AI Auditing Tools with Educators in Hawai'i. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, Article 217, 24 pages. doi:10.1145/3772318.3790958

[176] Tan Zhi-Xuan, Micah Carroll, Matija Franklin, and Hal Ashton. 2025. Beyond Preferences in AI Alignment. *Philosophical Studies* 182, 7 (2025), 1813–1863.

[177] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-Sensitive Algorithm Design: Method, Case Study, and Lessons. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 194 (Nov. 2018), 23 pages. doi:10.1145/3274463