Online Communities as Collaborative Testbeds for Collective Alignment

MATTHEW ZENT, University of Minnesota, USA
MALIK KHADAR, University of Minnesota, USA
SVETLANA YAROSH, University of Minnesota, USA
STEVIE CHANCELLOR, University of Minnesota, USA
HARMANPREET KAUR, University of Minnesota, USA

Representing human values in machine learning (ML) applications has received increased attention as end-users of real-world deployments surface new failure modes. Many technical approaches for preference alignment capture stakeholder values indirectly by reflecting end-user agreement with ML outputs, but face multiple validity concerns related to what they optimize for. Participatory ML (PML) has emerged as an alternative mitigation strategy, but it faces parallel challenges that undermine both its feasibility at scale and stakeholder empowerment goals. We first argue that online communities are well-equipped to improve the quality of human feedback for preference alignment. To achieve these benefits, we position online communities as collaborative testbeds where members serve as active partners rather than passive subjects. We recommend three foundations for collaborative testbeds to navigate core challenges in PML: 1) community artifacts as legitimate proxies for consultation, 2) situated deliberation as feedback, and 3) collective meaning cascades as signals for evolving preferences, or criteria drift. By leveraging communities' existing capacities to articulate, contest, and carry forward nuanced value considerations, our position provides a critical direction for sustainably applying PML principles at scales meaningful for technical ML work.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; • **Computing methodologies** → *Artificial intelligence*; *Machine learning*.

Additional Key Words and Phrases: online communities, machine learning, participatory machine learning, preference alignment

ACM Reference Format:
Matthew Zent, Malik Khadar, Svetlana Yarosh, Stevie Chancellor, and Harmanpreet Kaur. 2026. Online Communities as Collaborative Testbeds for Collective Alignment. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26), June 25–28, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/3805689.3812370

1 Introduction

The prevailing narrative of methods-driven Machine Learning (ML) calls for bigger datasets, models, and compute, leaving questions of efficacy and impact in real-world settings as an afterthought [125, 161]. Neglecting these use considerations during development and subsequent deployment of applications has resulted in accumulating evidence of new failure modes across domains [19] (e.g., healthcare [6, 29, 102, 160], education [114, 162], online health communities [34, 171], social media [167]). Beyond effectiveness, the potential for systemic harm to

Authors' Contact Information: Matthew Zent, zentx005@umn.edu, University of Minnesota, Minneapolis, USA; Malik Khadar, University of Minnesota, Minneapolis, USA; Svetlana Yarosh, University of Minnesota, Minneapolis, USA; Stevie Chancellor, University of Minnesota, Minneapolis, USA; Harmanpreet Kaur, University of Minnesota, Minneapolis, USA.

FAccT '26, Montreal, QC, Canada

ACM ISBN 979-8-4007-2596-8/2026/06
https://doi.org/10.1145/3805689.3812370

manifest in ML systems is well-documented, encompassing failures related to equity, interpersonal relationships, and society [15, 27, 47, 65, 103, 139, 150].

Preference alignment methods offer technical solutions for soliciting and including stakeholder input in the development and deployment of ML models (e.g., moral surveys, reinforcement learning from human feedback), but raise numerous questions of validity related to what they optimize for [48]. Reliance on surveys and crowdsourcing introduces context blindness, where decontextualized preferences fail to reflect the nuanced trade-offs and goals stakeholders navigate in actual decision-making environments [31, 45, 154]. These one-time elicitation strategies assume staticity, ignoring how preferences evolve over time as one's own understanding of system use and technological underpinnings change [21, 44]. Most critically, aggregating individual preferences risks preference collapse that rewards a statistical average inconsistent with any single stakeholder [11, 35]. At this point, resolving conflicting preferences becomes a problem of prioritizing stakeholder values. Addressing these validity concerns requires experimental settings where decision-making and collective deliberation are grounded in the contexts they affect.

Participatory ML (PML) has emerged as an alternative approach that centers stakeholder values in deployment environments, but it presents new challenges [101]. First, PML draws on participatory design's (PD) goals of stakeholder empowerment and fair design [63, 108], but the method's integration into contemporary ML practice often comes with extractive stakeholder participation when ML practitioners face pressure to generalize and adopt the latest models with limited control over realized benefits to stakeholders [24, 39, 144, 152]. Second, PML is often good at surfacing value tensions among participating stakeholders, but its aspirational fairness goals do not often translate beyond the research context. This challenge hinges on concerns over value illegitimacy when participants do not represent the broader stakeholder group, often because PML activities occur outside the contexts where collective decisions are normally made [18, 44]. Finally, limited organizational incentives for long-term stakeholder support and perceived lack of technical ambition introduce structural frictions that further undermine PML's commitments to stakeholders [101]. All of these challenges encourage practices that reduce PML's stakeholder empowerment goals to consulting for preference alignment, ultimately operating under the guise of participation (e.g., participation washing) [44, 144].

In this position paper, **we make the case for *online communities as collaborative testbeds*—environments where community members serve as active partners in model experimentation, debugging, and deployment rather than passive research subjects—to address fundamental challenges in representing stakeholder values in ML applications**. Online communities are social configurations of a shared virtual space for people to gather for information exchange, social support, learning, and entertainment [84]. They represent a situated context for applying PML principles at scales that are meaningful for technical ML work. We argue

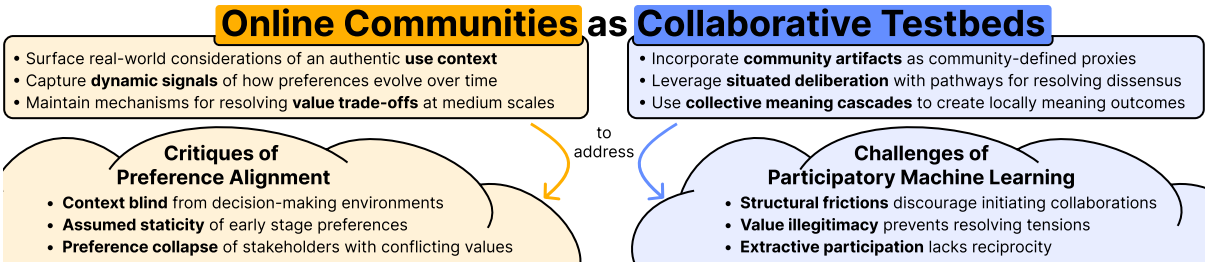

Fig. 1. Simplified overview of online communities as collaborative testbeds illustrating how properties of online communities address critiques of preference alignment and how qualities of our foundations for collaborative testbeds address challenges of PML. For example, collaborative testbeds incorporate community artifacts as community-defined proxies to address structural frictions that may discourage initiating collaborations.

that properties of online communities make them well-equipped to address the critiques of preference alignment. However, recent incidents also reveal the harms that result when members of online communities and their values are not centered (i.e., LLM deployment on r/changemyview [3]). Therefore, instead of simply treating them as sites for ML research, we argue for online communities to be collaborative testbeds where their members actively participate in shaping ML research and deployment, and describe concrete ways of achieving this vision (see Figure 1).

The remainder of the paper is structured as follows. We first define relevant concepts, and synthesize critiques of preference alignment and PML to establish the limitations of current approaches for representing human values in ML applications (Section 2). In Section 3, we describe three properties of online communities—use context, dynamic signals, and value trade-offs at medium scales—that equip them to address the critiques of preference alignment for ML. However, recognizing that this ideal case is challenging in practice, we recommend three methodological foundations of collaborative testbeds that transform online communities into tractable sites for tackling challenges of PML at scale (Section 4). Finally, we discuss open directions and alternative views to acknowledge barriers and clarify our position in order to help prioritize future ML work in online communities (Section 5). By synthesizing these critiques and methodological foundations into a forward-looking agenda, our goal is to motivate ML research to leverage online communities' ability to articulate, contest, and carry forward nuanced value considerations in ethical ways.

1.1 Defining Concepts

Aligned with prior work [84, 120, 166], we define *online communities* as spaces where members gather around a shared identity and set of norms for a variety of goals. We distinguish online communities from platforms, which provide infrastructure for people to gather, but do not necessarily foster the shared identity, established norms, or transparent and durable goals required to cultivate long-term and invested members. For example, ChatGPT users do not constitute an online community by this definition, but OpenAI's Developer Community[1] would. Similarly, the short-term, individualistic, and alienating environments of crowdworkers would typically not fit this criteria [26, 118]. In contrast, a group of virtual citizen scientists working toward a shared, fully visible, and enduring project like Foldit[2] would, even if their tasks mirror those in a micro-labor context. We discuss the attributes of online communities best suited for collaborative testbeds in Section 5.1.1 to invite empirical work across diverse communities.

By *participation*, we draw on Delgado et al. [44]'s conceptual framework, which articulates four modes of participation: consultation, inclusion, collaboration, and ownership. We also leverage this work's description of proxy participation and its risks, where stakeholders are ostensibly incorporated into ML decisions without direct involvement.

By *Machine Learning (ML)*, we refer to methods that learn or leverage predictive functions from data. While some related work uses the term AI or participatory AI, they often involve the ML pipeline. We consolidate these references under ML for continuity.

2 Related Work

To establish our motivation for online communities as collaborative testbeds, we first synthesize two research areas that support capturing human values in ML systems. We examine the contributions and limitations of these research areas, using **orange** to emphasize critiques of preference alignment methods and **blue** to highlight persistent challenges of PML.

[1]OpenAI's Developer Community is a Q&A forum for developers to collaborate, troubleshoot, and share experiential knowledge about OpenAI's APIs.

[2]Foldit is a crowdsourcing computer game that advances protein folding research, where the platform supports frequent community news and connections between users.

2.1 Preference Alignment

Preference alignment represents a shift in ML from engineered reward functions to explicit signals of human preference. The goal of preference alignment is to efficiently gather and learn reward functions or policies that result in end-user agreement with model outputs [83], typically using normative or performance criteria [48]. By normative preferences, we mean judgments that are more overtly value-laden. By performance preferences, we mean judgments about the correctness or quality of ML outputs. Critically, both are ultimately shaped by a combination of one's underlying values and use context [163]. Here, we review methods for incorporating preferences into ML, including moral surveys, reinforcement learning from human feedback (RLHF), active learning (AL), interactive machine learning (IML), and other human-in-the-loop (HiTL) approaches.

Recent years have seen significant advancements in RLHF and the development of large-scale preference benchmarks for alignment tasks. Early work in deep RLHF achieved high performance without access to a reward function using non-expert preferences on demonstrated behaviors [37]. Moral Machine datasets have been used in several projects to understand preferences for self-driving cars [12, 112]. More recently, LLMs have resulted in increased interest in alignment tasks through preference learning, spanning RLHF [31, 35, 74, 78], Direct Preference Optimization [121], Preference Ranking Optimization [146], personalization and persona testbeds [32, 174], and constitutional AI [14, 66].[3] These efforts have produced increasingly large preference datasets, with a growing effort on pluralism [147]. Kirk et al. [81] exemplified this agenda with PRISM, an alignment dataset with diverse preferences for cross-cultural general LLM use.

Beyond simple ranking, HiTL approaches enable richer forms of corrective [31] or cognitive feedback [105]. AL prioritizes strategic inclusion of experts using measures of diversity or uncertainty [5, 64, 107]—a popular strategy in communities on Zooniverse, a citizen science platform. IML is an iterative process of optimizing learning behavior [107, 122] that can produce useful policy steering vectors. Explainable AI (XAI) further reframes alignment to trust in outputs, emphasizing understandability, comprehensibility, and interpretability [10, 106]. While XAI methods have made significant strides, questions about how much and to whom remain unanswered [8, 43]. These methods are critical for effective backward alignment through assurance techniques and governance practices [70].

2.1.1 *Critiques of Preference Alignment.*

Despite these advancements, common preference alignment methods face significant validity concerns. These include *context-blindness*, *assumed staticity*, and *preference collapse*.

Context-blindness. Context-blindness refers to methods that derive abstracted preferences disconnected from realistic decision-making environments. This critique is particularly salient for moral preference alignment through surveys and crowdsourcing [31, 45, 154]. Preferences fundamentally depend on specific scenarios and use contexts [16]; as intended use cases of ML applications become increasingly general, users can struggle to provide meaningful feedback [101, 152]. Conversely, abstracting away algorithmic context by asking about human decision-making may fail to capture how preferences for algorithmic decision-making can differ from expectations held for people in equivalent situations [16]. Others have pointed to issues with certain alignment goals, such as the "helpful, honest, and harmless" paradigm, which remain unobjectionable because they are abstract [78, 79]. These concerns require methods that situate users and ML applications in contexts that surface concrete trade-offs and consequences.

Assumed staticity. Many alignment methods rely on one-time preference elicitation, introducing a different set of limitations by assuming preferences are static [21, 74, 76, 145, 154, 164]. Response instability between preference collection may reflect measurement problems introduced by context-blindness or psychological phenomena (i.e., fatigue, framing effects, or ordering bias) [74, 76]. Preferences also legitimately change as users' understanding of the system or context evolves [44, 75]. In the context of performance evaluation, Shankar et al. [137] coined the term "*criteria drift*" to describe the feedback cycle observed where evaluating model outputs refines users' criteria

[3] Supervised fine-tuning and pre-training are forms of LLM alignment. We choose not to discuss them in the context of preference alignment because these methods typically involve implicit and engineered measures of preference.

for evaluating outputs. This observation undermines methods that treat early-stage preferences as definitive. These dynamic patterns of preference highlight the need for methods that can distinguish legitimate change from measurement artifacts and for modeling approaches that can adapt to preference as a moving target.

Preference collapse. Perhaps the most persistent challenge for alignment methods like RLHF is aggregating across populations with conflicting preferences. Beyond majority bias, flattening diverse preferences into a statistical average can result in preference collapse where single-reward models converge to preferences inconsistent with any group of stakeholders [33, 130]. Conversely, over-personalizing to individuals introduces new risks that can inhibit collective progress [80, 123]. In response, pluralistic alignment proposes steerable personalization, distributional calibration to populations, and Overton pluralism (e.g., presenting many possible views) [147]. These approaches show promise, but raise important questions about what constitutes reasonable for different populations and use contexts [69, 147]. These questions, if answered poorly, can lead to over-personalization, marginalization, and presenting false balance. Scholars argue that pluralism is not about achieving consensus, but engaging in deliberative processes to find compromise while acknowledging power dynamics and value trade-offs [165]—dimensions largely absent from current preference alignment methods.

2.2 Participatory ML

Participatory approaches in ML trace their roots to Participatory Design (PD) in HCI, which emphasizes democratizing technology development by giving stakeholders genuine influence over design decisions [108]. PML's commitments to aligning algorithms with stakeholder values resonate with related design approaches, such as value-sensitive design and user-centered design. The interest in these approaches in ML is closely tied to the broader turn toward ML ethics, where PML is a key mechanism to advance algorithmic accountability to marginalized groups most affected by systemic ML failures.

The demonstrated successes in PML show how fully embedded stakeholders directly shape modeling decisions with situated knowledge to improve the quality of system outputs. ML development offers four entry points for participation: problem formulation, dataset development, model training and evaluation, and deployment and monitoring [39]. While many PML projects fail to meaningfully involve non-technical stakeholders [44, 48], exemplary cases demonstrate what deep engagement can achieve. Lee et al. [91] collaborated with a food donation service to design a model for equitable decision support by engaging stakeholders across the ML lifecycle, producing a system that outperformed historic allocation. Suresh et al. [151] worked with feminicide counterdata activists to conceptualize, build, and deploy an ML-based support system through iterative model refinement and evaluation. In peer production, long-standing collaborations on Wikipedia illustrate how editors curate data, build models, and audit outputs of edit quality classifiers, as seen in WikiBench [86], ORES on English Wikipedia [59], and article quality models on Dutch Wikipedia [60]. These projects exemplify the collaborative mode of the Parameters of Participation PAI framework, characterized by ongoing prototyping and shared decision-making [44]. However, as participatory work moves towards shared stakeholder ownership, the boundaries of PML become less clear. Community-initiated collaborations may not self-identify as PML and instead may surface as co-authored work [60, 141] or well-aligned partnerships [94]. This suggests that some successful community-aligned ML, particularly those focused on technical outcomes rather than collaborative processes, likely exist outside of recognized PML literature.

2.2.1 Challenges of Participatory ML. Much PML falls short of achieving higher modes of participation that enable collective exploration of ML systems and their impacts in contextual ways. As a result, a growing body of work highlights persistent challenges complicating PML's adoption. Some of these challenges are long-standing critiques of PD (*value illegitimacy*), while others were introduced with its adaptation to ML (*structural frictions* and *extractive participation*).

Structural frictions. PML faces structural frictions when practitioners and stakeholders negotiate deeply participatory collaborations. As participatory approaches shift decision-making power to communities, they may be perceived as limiting technical ambition, as participants can primarily articulate needs within their understanding of what is currently possible, which may or may not necessitate ML. This reflects a broader tension in ML research between application-driven approaches that engage with end users and methods-driven ML innovation [125]. A second friction arises from the increasing scale and opaqueness of modern models, which make it challenging to foresee impacts a priori [82]. This requirement for longitudinal testing [24] compounds the already limited organizational incentives to fully support ongoing, iterative participation across the ML lifecycle in light of prevailing scientific and financial pressures to pursue generalizability [101]. A third friction is that PML requires researchers to make initial assumptions about the problem domain to structure engagement with stakeholders [119]. Communities' desire for low-burden participation mechanisms when initial problem formulation is misaligned poses practical challenges for negotiating these preliminary considerations [97]. Together, these frictions have led ML practitioners to rely on proxy-based participation (e.g., convenience stand-ins and algorithmic models), which can introduce bias [44, 109].

Value illegitimacy. We adapt and modify limitations of the prototypical PD workshop to coin the term value illegitimacy [18, 63]. Similar to issues with convenience stand-ins, this challenge stems from persistent questions of legitimacy regarding who is invited to participate, who is able or willing to participate, and who ultimately represents a community [63]. This is a primary conceptual concern with PML methods that lack mechanisms or the authority to resolve complex value tensions outside of the research context. Because of these gaps, codifying values and norms is more appropriately shaped through collective democratic processes and community governance [18, 42, 151]. Value conflicts remain a persistent challenge in community-centered and participatory-inspired methods (e.g., STELA [17]). Without a clear path toward resolving the conflicting needs of different groups, participatory efforts risk being performative [40, 144]. These challenges intensify at global, cross-cultural scales, where neglecting specific groups can risk advances that reinforce harms in historically exploited communities [9, 170]. These contexts also raise new questions about the conditions under which people will speak out [85, 153]. Finally, as ML systems grow in size and scope, the number of affected stakeholders increases, further complicating questions of representation and who should have a voice in participatory processes [7].

Extractive participation. Contemporary ML development also hinders meaningful participation from members. To this end, Sloane et al. [144] argue PML is "*necessarily extractive at the level of ML design, development, and deployment*" rooted in the desire for ML systems to generalize beyond the context of participation [104]. Scholars have also questioned whether participation is equally appropriate at all stages of the ML lifecycle [48]. Increasingly opaque models manifest delayed and unpredictable downstream impacts, complicating the realization of reciprocal benefits to justify the labor of participation. Additionally, the rapid release cycles of foundation models can render participation obsolete or impose a participatory ceiling on what researchers can meaningfully provide to stakeholders [152]. Considering PML's ethical commitments, the substantial labor needed to sustain meaningful participation often disproportionally burdens marginalized groups who participate out of a desire to be heard [172]. Similar to others [44], we conclude that PML needs to focus on measurable outcomes that meaningfully reflect stakeholder contributions.

3 Why Online Communities?

Having examined the limitations of current preference alignment methods, we advance our position by describing how online communities help address key concerns discussed in Section 2.1.1. We identify the benefits that online communities offer for modeling human preferences, establishing them as promising testbeds for ML experimentation, debugging, and deployment.

3.1 Use Context

Online communities offer insights for the ML lifecycle that can only emerge from authentic use contexts. The insights gathered from online community members address the **context-blindness** critique of ML alignment methods by reflecting the goals of stakeholders with a genuine stake in the health of their community.

To illustrate how members' sustained engagement within their communities' use context enables them to articulate nuanced corrections, consider this hypothetical example inspired by Chimp&See:[4]

> *Researchers developing a computer vision model introduce it to the community as an "AI assistant" that predicts species and provides heatmap explanations to accelerate the members' annotations. Members' investment in producing high-quality labels and their accumulated experience enable them to catch systematic errors in explanations while still benefiting from faster initial predictions. For misclassified images, members contrast the assistant's explanations with their own sets of heuristics, revealing both why the model failed and what features it should have prioritized. This situated feedback enables researchers to identify dataset biases (e.g., species/habitat correlates) and refine the model's attention mechanisms.*

As demonstrated by members' ability to contrast model behavior with their personal experiences, online communities provide rich corrective feedback that is difficult to replicate in decontextualized crowdsourcing. Members' situated expertise developed through contributing to their community's goals enables them to provide specific and actionable feedback [176]. As illustrated in the example above, these benefits are amplified through eXplainable AI (XAI). Feedback in XAI is often limited to high-risk domains with experts due to the investment and expertise required to meaningfully engage with explanations (e.g., in healthcare [129]). Online communities offer a distinct opportunity for XAI to reach broader audiences of invested users [23, 43].

A key benefit of authentic use contexts arises for problem formulation and deployment, where members are invested in the real-world considerations of their community. Anticipating deployment issues is difficult, and is made even more difficult with the rise of foundation models where the intended use contexts can be unclear or too broad [152]. Engaging with communities that have established goals and concrete use contexts enables ML practitioners to observe preferences and behaviors grounded in actual decision-making environments, overcoming the confounds introduced by hypothetical scenarios and decontextualized assessments [31, 152]. Furthermore, members' investment in outcomes improves performance for validation tasks as they are less susceptible to fatigue and biased heuristics than crowdworkers [31, 126].

3.2 Dynamic Signals

Online communities have dynamic signals that capture how preferences and system efficacy evolve through positive and negative feedback loops introduced by deployed systems [41]. Members alter their behaviors as they learn about system capabilities and discover failure modes. Community traces and version-controlled wikis reflect shifts in system performance and user expectations that mitigate concerns over **assumed staticity** in one-time elicitation methods.

To see how situated knowledge of a community's evolving preferences mitigates failures introduced by assuming static preferences, consider this example adapted from prior work [92] to a highly moderated Q&A subreddit community:

> *Researchers deploy drift detection to a community's existing auto-moderation tools. Data shifts from first-time posters' language use trigger periodic alerts corresponding to large increases in auto-removed posts. The remediation strategy for these spikes is unclear from metrics alone. They could indicate correctly filtering genuinely unwanted content (requiring no action) or incorrectly removing legitimate*

[4]Chimp&See is a citizen science community where members identify which species are contained in images of wildlife. Annotations require significant lay-expertise due to the visual similarities between closely related species.

contributions (requiring retraining to prevent newcomer attrition). Researchers consult community meta-discussion channels, where members reveal two distinct observations causing increased newcomers: community features on Reddit's front page (transient and unhealthy) and cross-posts on topically similar subreddits (potential growth). Researchers incorporate the traffic source patterns identified by members into the drift detection algorithm, enabling higher quality monitoring in the absence of ground truth data to confirm predictive accuracy.

This example shows the value of members' dynamic understanding of community health, where assuming static preference would treat drift as a binary signal requiring retraining. Not all drift represents failures, and distinguishing harmful from benign drift requires contextual knowledge [92, 136, 140]. Online communities enable experts to emerge based on their contextual knowledge within a community, affording new directions for deployment monitoring.

Communities surface dynamic signals to inform critical decisions across the ML lifecycle. Dynamic feedback loops transform members' capacity to provide valuable preference signals. What distinguishes community-based feedback is members' situated understanding of community goals, enabling reflectivity as they observe changes introduced by ML deployments [38]. This parallels Interactive ML approaches, where repeated engagement within and between research enables communities to progressively articulate more sophisticated feedback to update modeling decisions [31, 122]. Members' evolving understanding of community health and goals often becomes formalized through versioned community artifacts—updated rules and wikis that capture legitimate change. Unlike static preference datasets [87, 149], these artifacts record changing preferences and conceptual drift. This enables the construction of valid community benchmarks where evaluation criteria evolve with the community to combat errors from sample selection bias and nonstationarity [57, 111].

3.3 Value Trade-offs at a Medium Scale

Online communities occupy a liminal space between individuals and broader society, providing natural settings for negotiating value trade-offs that cannot be resolved through individual preference aggregation. This medium scale ranges from hundreds to hundreds of thousands of members and addresses the **preference collapse** critique by offering a context pluralistic enough to surface genuine value tensions while maintaining mechanisms for their resolution.

Recent work has formalized the computational intractability of preference alignment at a societal scale, calling instead for constraining representation goals, scoping to realistic contexts, or accepting super-polynomial computational costs [130]. Online communities naturally decompose this alignment problem into subproblems by scoping to considerations of their unique use context. Additionally, members share enough context and goals to establish common ground to negotiate underlying values. Unlike individual-level personalization that risks echo chambers [142], online communities surface diverse perspectives by maintaining the feasibility of meaningful deliberation when misalignment occurs. Deliberative processes also help avoid exploiting short-term satisfaction of individuals at the expense of holistic community goals [80, 123, 138]. The medium scale supports individuals feeling like their voices are part of the conversation and allows for collective sensemaking by situated experts—members able to articulate value trade-offs with legitimate authority (e.g., moderators on subreddits, experienced Wikipedia editors). Beyond making alignment more tractable, these communities more effectively cultivate group identities [67]. In turn, this helps motivate the participation required to sustain deliberation in online communities [88].

To illustrate how value trade-offs in online communities can be useful for resolving preference collapse, consider the following hypothetical case inspired by content moderation challenges in mental health support communities:

Researchers developing a content moderation model observe members' frustration with their community's existing topic-based moderation tool. Comments on removed posts reveal competing values: removing potentially harmful content (safety) versus preserving authentic peer experiences (support and relatedness), which collapses into either over-removal (prioritizing safety at the expense of authentic connection), under-removal (prioritizing authenticity while permitting harm), or a mechanical failure (sterilized peer support). A meta-post for members to deliberate on this issue reveals that while members do not agree on what topics constitute harm, they have negotiated a nuanced position clarifying this decision boundary: "warn, don't remove" for posts describing past struggles, but immediate removal for content encouraging current harmful behaviors. This distinction only becomes tractable because community members share an understanding of what constitutes "support" in their community. Researchers then incorporate a separate model to detect past vs. present tense content to reflect actual community values rather than an abstracted safety vs. expression trade-off.

This example demonstrates how deliberative opportunities in online communities can inform decisions throughout the ML lifecycle. Online communities surface value tensions invisible in decontextualized surveys but critical for deployment. The visible negotiation of trade-offs provides ground truth about legitimate value pluralism versus measurement noise in human feedback. For model training, their medium scale enables strongly specified target populations for distributionally or Overton pluralistic models [147]. During deployment, communities demonstrate mechanisms for managing ongoing value misalignment without requiring impossible consensus. They also offer potential for situated expertise to provide different perspectives on deployments.

Importantly, the example shows that online communities do not always reach consensus on value trade-offs—and that is a feature, not a bug. Instead, deliberation aimed at compromise is preferable in pluralistic democracies so long as it does not sacrifice members' foundational values [165]. In this way, online communities develop social and technical mechanisms to make decisions that further shared goals despite disagreement (e.g., voting systems, moderator discretion, A/B testing of rule changes, and meta-discussions to reach a state of productive dissensus [101]). These kinds of collective sensemaking processes are well-documented in online community research. In large, radically open communities like Wikipedia, members coordinate around mutually relevant tasks and often commit to partial but acceptable interpretations of policy, knowledge, and norms [110]. At medium scales, more private communities like TuDiabetes often value pluralistic opinions over consensus to negotiate shared meaning [98].[5] These practices themselves can reveal how to design ML systems that accommodate value pluralism. Online communities illustrate which disagreements are resolvable through clarifying shared understanding versus which reflect genuine value diversity that systems must navigate [89].

4 Online Communities as Collaborative Testbeds

As described above, online communities offer valuable testbeds for ML researchers. However, recent incidents reveal the harms resulting from prioritizing research goals over community values (e.g., Linux hypocrite commits [49] and the r/changemyview incident [3]). Therefore, our position reframes online communities as *collaborative testbeds*—environments for model experimentation, debugging, and deployment that prioritize collaboration and ownership. Online communities are increasingly developing practices aligned with this vision. Here, we discuss three methodological foundations to leverage these capacities in future ML research and help address persistent challenges in PML (Section 2.2.1).

[5]TuDiabetes, now Beyond Type 1, is an online health community for people living with diabetes, and their friends and family, to give and exchange peer support.

4.1 Community Artifacts as Legitimate Proxies for Consultation

Online communities maintain wikis, guides, and rules that can serve as legitimate proxies for consultation in collaborative testbeds. These community artifacts help mitigate the **structural frictions** of PML by offering low-burden mechanisms to navigate early modeling considerations with community-defined priorities. While not all community artifacts emerge through democratic processes, when accessible to members, they represent negotiated value trade-offs that encode collective priorities. In some communities, these artifacts gain legitimacy through members' ability to contest them. For others, legitimacy is mediated through stewardship where trusted guardians protect community goals [172].

Most online communities produce negotiated artifacts, and we observe a trend that these are increasingly articulating values related to research and ML. Wikis directed at newcomers document collective knowledge and set expectations for normative behavior, providing researchers with essential context for understanding deployment environments and community priorities before engaging in system design. r/AskHistorians and r/ChangeMyView recently published research guidelines that prioritize community values; specify data scraping permissions; and outline privacy expectations, effectively encoding consent [58, 132]. Communities on various platforms (e.g., Stack Exchange, Discord, Reddit, or Zooniverse) afford members channels to deliberate about feature requests and collective goals. These meta-discussions explicitly describe trade-offs and evolving priorities for researchers seeking to better understand community values. Finally, community-curated datasets and data donations that are more prevalent in peer production projects and data-oriented communities formalize legitimacy through acceptable use statements, access requests, and member transparency.

Community artifacts offer concrete utility for problem formulation, iterative model development, and evaluation criteria. For instance, community feature requests and resulting deliberations can directly align problem formulation with real needs, whereas wikis and guidelines articulate evaluative criteria through examples and justifications of community priorities. While these artifacts are inevitably under-specified for many research objectives, they initialize design decisions closer to deployment realities [143], making subsequent higher-fidelity engagement with communities more productive. To demonstrate this utility, consider the following hypothetical example in r/changemyview:[6]

> *Researchers design a generative argument quality assessment model (GAQAM) for r/changemyview, leveraging the community's extensive wiki documenting the dos and don'ts of effective argumentation for policy-based steering (i.e., positive and negative examples of effective ad hominems). Similar to steering for chain-of-thought (COT) compression [13], this approach rewards reasoning in the community's shared language. In offline evaluations, the initial policy-constrained model achieved only marginal increases in predictive accuracy compared to an unbounded COT baseline. Greater benefits emerge during deployment, where starting from members' collective understanding enables richer corrective feedback at intermediate reasoning steps to directly update the steering vector, resulting in iterative performance gains.*

Critically, these legitimate proxies provide researchers access to well-understood community values without requiring direct member consultation. This enables researchers to initialize their understanding of the problem domain and align preliminary design decisions with the community before engaging members. This is valuable for reducing members' burden of educating researchers or participating in exploratory work that may not yield valuable outcomes for communities [97, 172]. Moreover, when researchers articulate the normative rigor of technical choices using the documented values of an online community, they earn new forms of epistemic authority for social claims [82, 113].

[6] r/changemyview is a large deliberative subreddit where users invite others to challenge their views through structured, good-faith argumentation.

4.2 Situated Deliberation as Feedback

Several online communities engage in deliberation about governance and decision-making [53, 169]. Deliberation is a process where people "*carefully examine a problem and arrive at a well-reasoned solution after a period of inclusive, respectful consideration of diverse points of view*" [56]. Situated deliberation addresses **value illegitimacy** by reflecting real decision-making environments as members enact various roles in their community to negotiate value trade-offs. By following community participation norms, this feedback on ML interventions, outputs, and design considerations provides a legitimate pathway towards adjudicating conflicting needs.

Online communities frequently engage in governance deliberation and offer opportunities for researchers to participate in discussions. Virtual spaces allow vast audiences to share their perspectives, making them widely recognized for their democratic potential [116, 134]. The form and function of deliberation varies by community, from conflict resolution on Wikipedia [20, 68, 155] to decision-making processes of open-source software [73, 156], to policy experimentation in subreddits [100]. Given rising controversy surrounding ML applications [99], many communities are actively deliberating about generative models' impact on their health and values (i.e., iNaturalist [168],[7] StackOverflow [2], and Wikipedia [4]). Meta-discussions are often acceptable community spaces for researchers to participate in or structure deliberation with organizers' approval—further supported by moderation to ensure civil and inclusive discussions.

Building on these established practices, collaborative testbeds enable situated deliberation to guide both model evaluation and iteration. This facilitates more robust feedback about the practical utility, anticipated impact, and acceptance of ML applications [46, 154]. Discussions represent the intermediate sentiments of members, creating a space to jointly contest ML outputs. Forward-facing deliberation provides rich corrective feedback on stale or biased models rather than being constrained to a limited set of alternative solutions [124]. To illustrate how deliberation can inform iterative model development, consider this example inspired by Project Sidewalk [95]:[8]

> *Researchers introduce an "AI assistant" to automatically identify issues and assign severity ratings, which citizen scientists can validate. Some members consistently adjust the severity ratings for certain types of issues (i.e., uneven sidewalks), which prompts researchers to create a meta-thread to discuss annotation disagreement. By discussing examples, members describe how their experiences with different types of mobility devices shape their assessments. Researchers introduce the idea of adding multiple severity ratings to the AI assistant based on mobility context, prompting further deliberation about how to present or filter multi-dimensional predictions to future annotators.*

Beyond iterating through feedback, the deliberative process offers value legitimacy for resulting decisions. When highly visible, it helps foster a sense of shared stewardship among stakeholders, which can increase buy-in from members on resulting actions. Community meta-discussions often reveal genuine uncertainty about evolving priorities (e.g., equipoise [101]), which can incentivize communities to commit to empirical work to inform decision-making. Deliberation on specific model outputs also improves annotation quality, as discussion about reasoning helps members converge on shared understanding [77, 133]. Effective deliberation also creates space for dissensus, which is valuable when people with different values engage in an evaluation task [89]. Iterative feedback and productive uncertainty position situated deliberation as a mechanism to strengthen value alignment in collaborative testbeds.

4.3 Collective Meaning Cascades as Signals for Criteria Drift

Online communities pursue and refine collective goals (e.g., maintaining quality discourse or supporting member growth) through evolving artifacts that encode collective understanding. Halfaker et al. [60] coined the term *collective meaning cascades*, whereby inherently imperfect models translate and modify the meaning of information they

[7] iNaturalist is a citizen science community for members to contribute data to biodiversity research.
[8] Project Sidewalk is a citizen science community focused on identifying pedestrian accessibility issues from street view data

represent, a phenomenon similar to criteria drift [90, 137]. In online communities, criteria drift describes how members' understanding of what a correct, useful, or good model output is changes as they interact with and evaluate ML systems in their community. By anchoring this refinement in community practice, collective meaning cascades help reconcile **extractive participation** in PML by simultaneously producing locally meaningful outcomes. In collaborative testbeds, iterative refinement of community artifacts is important for sustaining community participation across multiple ML research projects and for fostering localized understanding [44, 172]. This is valuable when researchers have limited control over the underlying algorithms or no intent to support applications long-term.

Many online communities already iteratively refine their goals and shared artifacts, and we occasionally find that researchers play a part in this process for better or for worse. Communities frequently update their collective understanding through rule changes in response to events, edge cases, and breaches of trust from researchers [93]. Across platforms, communities are actively grappling with rules for GenAI use in relation to their goals for authenticity, quality, and community values (i.e., subreddit communities [96], StackOverflow [22], and Wikipedia [52]).

Beyond reactive changes, we argue that observing collective meaning materialize in shared community artifacts is a healthy signal of *higher* modes of participation made possible by well-aligned problem formulation. For example, members annotating community data cascading to revised labeling guides and definitions on citizen science projects [30] and Wikipedia [60]; or members curating datasets cascading to acceptable use statements for community data [86]. Community members are often interested in their collective values and goals [172], and some communities have implemented their own mechanisms to embrace this curiosity; r/AskHistorians collaborates with Cornell to distribute census surveys about participation motivations, and r/changemyview created r/ideasforcmv for improving the community. These observations motivate researchers to align model evaluation criteria with communities' unique definitions of what makes them better in order to support both research and community self-knowledge [166].

Refining community-situated knowledge benefits the ML pipeline by providing use-informed incentives for iterative model development and evaluation. As with under-specification in LLM reward modeling [148], this provides opportunities for participation to mediate meaning cascades, which in turn refine under- and misspecified modeling requirements. Community artifacts inherently reflect a point in time, and while many are forward-looking by considering broader notions of community health, they can fall victim to stale modeling. This presents an opportunity for members to meaningfully engage with model outputs, serving a dual purpose of auditing collective understanding and modeling assumptions. For modeling, these updates can directly produce high-value labels, feedback to add or prune features, or new deployment or acceptability considerations. We revisit our hypothetical r/changemyview scenario to illustrate how initializing with community artifacts enables collaborative meaning cascades:

> *Again, researchers are designing a policy-based GAQAM, but want to add audience awareness capacities—tacit knowledge for experienced members, but absent in shared artifacts. Testing their GAQAM on offline benchmarks using historic community deltas (awarded when the OP changes their view) reveals discrepancies, but it is unclear whether the performance degradation is due to latent personalization strategies (add to wiki), stale preferences (update wiki), or encoded bias (actions vary). Researchers analyze the residual to identify high-importance text features and present exemplars of each strategy to members in a meta-post. The process of reflecting on exemplars refines members' understanding of effective uses of appeals to authority, popularity, and emotion, which are added to the wiki and update the researchers' policy vector.*

This wiki-informed steering approach enables researchers to re-encode policy vectors rather than fine-tune models, which substantially reduces the computational costs of iteratively maintaining alignment. Rather than prescribe a single approach, collaborative testbeds ask researchers to consider how different alignment methods integrate with communities' existing artifacts, practices, and norms (e.g., community-led artifacts such as

policy steering [36], constitutional AI [66], or curated datasets [86]). Collective meaning cascades aim to bridge contextual gaps in applied ML by establishing increased construct validity in evaluation criteria, shifting the focus from accuracy to use-informed utility [34, 60].

5 Discussion

5.1 Open Challenges and Limitations

While reframing online communities as collaborative testbeds offers a promising direction for actualizing PML principles at scale, they face new practical and ethical barriers. Here, we identify key limitations in our position to help prioritize future alignment research with online communities.

5.1.1 Not all communities are suitable. Communities vary in their organizational maturity, which shapes the effort and resources required to establish community artifacts as legitimate proxies for participatory consultation. Community artifacts will often be underspecified for research objectives. We argue that this friction will diminish over time within the virtuous cycle of collaborative testbeds, as illustrated by more mature communities (i.e., Wikipedia, Zooniverse, InTheRooms,[9] and Stack Exchange sites). These challenges highlight one direction for future research on mechanisms that build community capacity and translate resources between similar communities (e.g., templates for research guidelines or privacy-preserving tools for community data) to lower the barrier to entry for collaborative testbeds.

The efficacy of situated deliberation as feedback depends on how community members choose to engage with it or not. While dissensus can surface value trade-offs, resulting actions can fragment communities, similar to how new rules can drive user abandonment [41]. The added pressure in community discussions can further alienate those who routinely participate less (i.e., lurkers or low contributors) [131]. Conversely, the need to keep community-situated processes open and accessible invites the risk of bad actors and data poisoning, forcing researchers and likely implicating community organizers to grapple with the complex problem of weighing contributions. In light of these deliberative challenges, collaborative testbeds offer valuable opportunities to build our understanding of how platform designs afford deliberation and grow the limited body of work examining the outcomes of deliberative processes in practice [54].

For other communities, these barriers may make collaborative testbeds inappropriate. Communities that lack meaningful collaboration between members or with limited affordances or norms for discussion undermine the benefits of deliberation components of collaborative testbeds. Echoing Section 1.1, others may lack transparent and durable goals that cultivate situated expertise for weighing value-tradeoffs with legitimate authority. In such cases, community artifacts may also lack legitimacy. In other ways, research burnout has led some communities to establish rules against specific research activities (e.g., surveys on subreddits [25, 117]). This underscores why collaborative testbeds must move beyond online communities as a recruitment opportunity. Truly collaborative testbeds give communities the power to decline. Communities that perceive high risks or have suffered negative research experiences in the past may decide that potential benefits do not justify research engagement.

While criteria for suitable communities remain an open question, Computer-Supported Cooperative Work (CSCW) scholarship on participation offers promising directions for future work across two dimensions: motivation and collaboration. First, we consider members' motivations to participate. Communities where members are driven by a sense of perceived impact are better positioned to support the member-driven benefits of a situated use context [88, 115]. For example, those with strong group identities that value pluralism over consensus are particularly good starting points [67, 98]. Members' motivations are not only important for achieving the benefits

[9]InTheRooms is an online recovery community with paid opportunities to conduct health and social computing research within the community.

of value trade-offs, but research with these communities is more likely to produce novel and valuable contributions for their members and academic scholarship. Second, we consider how members collaborate together. It is important that members' individual participation depends on the contributions of others to develop a collective understanding (e.g., the strong remix norms of transformative fandom communities [50]). These collaborative communities may also desire more control over ML for socially embedded practices [128]. And, they often set expectations for data re-use with members. These norms support increased data provenance, tool provenance, and transparency [173]. In doing so, they naturally produce the kinds of artifacts that orient new members and reveal dynamic signals for research. Members' motivation to collaborate helps identify communities where establishing collaborative test beds is most likely to be mutually beneficial.

5.1.2 Not all research questions are suitable. Prioritizing mutual benefit narrows the scope of research questions that can be pursued within a given online community. Research alignment is a long-standing challenge in community-based participatory research [97]. Community-driven lines of inquiry are likely to produce unfamiliar benchmarks, have limited broader applicability, and result in simple solutions that may outperform SOTA models in situated settings. Others argue these are important complements to methods-driven ML [125]. We contend that as research engages more deeply with communities, these questions will become increasingly sophisticated. Drawing on community-engaged research practices, collaborative testbeds demand give-and-take relationships where highly aligned research questions build capacity to articulate considerations valuable for less aligned work. Moreover, researchers have the advantage of drawing on the diverse landscape of online communities with different interests and goals. For long-term relationships, researchers might strategically pair community-prioritized studies with inquiries of wider scientific importance.

Beyond reciprocity constraints, online communities impose methodological restrictions for research [172]. For instance, deception-based studies are typically impermissible in online communities.[10] This reality places additional demands on the normative, reporting, and interpretive rigor of ML [113], requiring that decisions be justified not only to peers but also to community members. This challenge mirrors ongoing tensions between application-driven and methods-centric ML [125], and is essential for closing the claim-reality gap between stated research goals and actual impact [82]. For ML agendas incompatible with existing communities, others have proposed creating research communities. MovieLens is a standout example [61], but others have proposed drawing on open-source and peer production models for centering research activity as a core community goal in RecSys [28] and AI feedback [45].

5.2 Alternative Views on Collaborative Testbeds

Collaborative testbeds present a vision for how ML research should engage with online communities, but this approach may face reasonable objections. Here, we clarify our position to guide thoughtful adoption.

5.2.1 Online communities are not an important use context for ML. Some may view online communities as peripheral to ML innovation because they represent niche domains, or because use context considerations appear orthogonal to methodological advancement [125, 175]. This view overlooks online communities' high representation in LLM training data.[11] Researchers face mounting barriers to access these ostensibly open resources [1, 157], making these considerations increasingly difficult to avoid. Our position may extend beyond online communities to contexts with similar properties; for example, workplaces have collective goals, dynamic signals, and the capacity to articulate value trade-offs, suggesting value in artifacts, deliberative feedback, and meaning cascades to support sustainable ML in CSCW [62]. Others may contend that use-context concerns fall outside the scope of their

[10]Turing tests are diminishing in scientific value [71]. Instead, consider the online community-aligned research direction of detecting algorithmically generated content.

[11]For example, Wikipedia and StackExchange comprised ~6.5% of pre-training data in Llama [158]

ML work. In response, many argue that social claims about applicability and impact must incorporate deployment considerations throughout development rather than deferring them indefinitely [29, 34, 72, 82, 113, 125]—online communities present a tractable context to pursue this goal.

5.2.2 Collaborative testbeds introduce logistical challenges. This view raises open questions about institutionalizing insights from online communities. We agree that our emphasis on grounded domains and specific use cases for ML introduces new challenges for integrating feedback into larger base systems. Others have called for collective action in different environments and by different groups to eventually represent a larger stake in the industry to impact general-purpose models [135, 150]. Still, reconciling insights across collaborative testbeds remains challenging. Not only because of conflicting preferences, but communities may also maintain fundamentally different technical representations of those preferences (e.g., one may use a constitution [66] while another uses examples for policy steering [36]). Collaborative testbeds will produce artifacts that touch different stages of the ML lifecycle, some more portable or complete than others. Indeed, future work should examine how and when different methods integrate smoothly with both community practice and post-training pipelines in real-world deployments.

5.2.3 Collaborative testbeds do not address the fairness goals of PML. Under this view, collaborative testbeds fall short of achieving fairer outcomes because they do not address biases within communities themselves or among those who engage in research activities. We contend that thoughtfully embedding research within community structures helps address the latter, but deliberative processes may amplify the voices of organizers and active contributors, excluding lurkers and peripheral members [55]. Ultimately, there is no lightweight solution for participation as justice, which requires long-term partnerships with mechanisms for recourse to address structural change [144]. Rather than claiming to resolve bias concerns, collaborative testbeds foreground specific community-defined fairness outcomes [44].[12] However, this does not diminish the importance of reflexivity about representation and power within online communities.

5.2.4 Collaborative testbeds give research an excuse not to pay participants. Some may object that collaborative testbeds mask exploitation by claiming intrinsic motivation or community benefit. In this view, assuming members' motivations risks overlooking invalid assumptions about what research offers the community, heightened by contemporary ML practices [144]. While intrinsically motivated participation can produce high-quality outcomes [127, 159], we largely agree with this perspective. We clarify that participation is work [144], and online communities as collaborative testbeds do not absolve researchers from providing financial compensation when appropriate; however, we expand on what *compensating a community* could entail. After research engagement, researchers offer additional incentives to members who choose to apply insights by updating, refining, or creating shared community resources that benefit the collective.

6 Conclusion

By positioning online communities as collaborative testbeds, we address fundamental challenges in representing human values in ML applications. Our argument demonstrates how online communities offer a valuable context with dynamic signals of preferences, grounded in a use context, with mechanisms for resolving value trade-offs at a medium scale. We illustrate how collaborative testbeds address core challenges of structural frictions, value illegitimacy, and extractive participation in PML by foregrounding online communities' existing capacities for generating community artifacts, engaging in situated deliberation, and refining their own practices. By motivating research to leverage community infrastructures for articulating, contesting, and carrying forward nuanced value considerations across projects, these foundations provide a critical direction for sustainably applying PML principles at scales meaningful for technical ML work.

[12]For example, gender equity on Wikipedia or topic diversity on r/AskHistorians. We note that not all communities have fairness goals. These may fall under *Not all communities are suitable* for collaborative testbeds if researcher-community values are fundamentally misaligned.

7 Endmatter

Generative AI Usage. Claud Sonnet 4.5 was used to edit grammar and styling for text clarity and conciseness. No generative AI was used to create figures or images in this paper. All generative AI use was carefully reviewed by the authors and represents our original ideas and concepts.

Ethical Considerations. While we discuss the ethical barriers and potential adverse impacts of adopting online communities as collaborative testbeds in Section 5, we elaborate on additional ethical considerations of this work here. We intentionally avoid relying solely on ethical arguments to justify our position and instead focus on validity and modeling benefits to reach a broader audience. We followed ethical recommendations for online community research to prepare this position paper [51, 172]. Because we draw on extensive examples from r/changemyview, we follow the community's guidelines for moderator review before initiating this work. Specifically, we communicated the intentions of this position paper, clarified aspects that do or do not support their objectives, shared hypothetical examples, and avoided including specific member quotes.

Acknowledgments. We are grateful to the moderators at r/changemyview for their interest in this work. We thank Aaron Halfaker for his guidance on participatory engagement with online communities. This work was supported by NSF ER2 Award #2220509.

References

[1] 2023. Data API Terms. https://redditinc.com/policies/data-api-terms Publication Title: Reddit Inc Homepage.

[2] 2023. New blog post from our CEO Prashanth: Community is the future of ai. https://meta.stackexchange.com/questions/388401/new-blog-post-from-our-ceo-prashanth-community-is-the-future-of-ai Publication Title: Meta Stack Exchange.

[3] 2025. Meta: Unauthorized experiment on CMV involving ai-generated comments. https://www.reddit.com/r/changemyview/comments/1k8b2hj/meta_unauthorized_experiment_on_cmv_involving/

[4] 2025. Wikipedia:Village pump (policy)/Good faith and AI-generated comments. https://en.wikipedia.org/wiki/Wikipedia:Village_pump_(policy)/Good_faith_and_AI-generated_comments Publication Title: Wikipedia.

[5] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3173574.3174156 event-place: Montreal QC, Canada.

[6] Molla Imaduddin Ahmed, Brendan Spooner, John Isherwood, Mark Lane, Emma Orrock, and Ashley Dennison. 2023. A systematic review of the barriers to the implementation of artificial intelligence in healthcare. Cureus 15, 10 (2023). Publisher: Cureus.

[7] Leah Hope Ajmani, Nuredin Ali Abdelkadir, and Stevie Chancellor. 2025. Secondary Stakeholders in AI: Fighting for, Brokering, and Navigating Agency. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25). Association for Computing Machinery, New York, NY, USA, 1095–1107. doi:10.1145/3715275.3732071

[8] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information fusion 99 (2023), 101805. Publisher: Elsevier.

[9] Payal Arora. 2025. Creative data justice: a decolonial and indigenous framework to assess creativity and artificial intelligence. Information, Communication & Society 28, 13 (2025), 2231–2247. arXiv:https://doi.org/10.1080/1369118X.2024.2420041 doi:10.1080/1369118X.2024.2420041

[10] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58 (2020), 82–115. doi:10.1016/j.inffus.2019.12.012

[11] Anne Arzberger, Stefan Buijsman, Maria Luce Lupetti, Alessandro Bozzon, and Jie Yang. 2025. Nothing Comes without Its World - Practical Challenges of Aligning LLMs to Situated Human Values through RLHF. In Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society (AIES '24). AAAI Press, San Jose, California, USA, 61–73.

[12] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The Moral Machine experiment. Nature 563, 7729 (Nov. 2018), 59–64. doi:10.1038/s41586-018-0637-6

[13] Seyedarmin Azizi, Erfan Baghaei Potraghloo, and Massoud Pedram. 2025. Activation steering for chain-of-thought compression. arXiv preprint arXiv:2507.04742 (2025).

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

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

[16] Chandrayee Basu, Qian Yang, David Hungerman, Mukesh Singhal, and Anca D. Dragan. 2017. Do You Want Your Autonomous Car To Drive Like You?. In Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI '17). Association for Computing Machinery, New York, NY, USA, 417–425. doi:10.1145/2909824.3020250 event-place: Vienna, Austria.

[17] 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.

[18] 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 (EAAMO '22). Association for Computing Machinery, New York, NY, USA. doi:10.1145/3551624.3555290 event-place: Arlington, VA, USA.

[19] Arno Blaas, Priya Ronald DCosta, Fan Feng, Andreas Kriegler, Zhaoying Pan, Tobias Uelwer, Jennifer Williams, Yubin Xie, and Rui Yang. 2025. I Can't Believe It's Not Better: Challenges in Applied Deep Learning. In ICLR 2025 Workshop Proposals.

[20] Laura W. Black, Howard T. Welser, Dan Cosley, and Jocelyn M. DeGroot. 2011. Self-Governance Through Group Discussion in Wikipedia: Measuring Deliberation in Online Groups. Small Group Research 42, 5 (2011), 595–634. arXiv:https://doi.org/10.1177/1046496411406137 doi:10.1177/1046496411406137

[21] Kyle Boerstler, Vijay Keswani, Lok Chan, Jana Schaich Borg, Vincent Conitzer, Hoda Heidari, and Walter Sinnott-Armstrong. 2024. On the stability of moral preferences: A problem with computational elicitation methods. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, Vol. 7. 156–167.

[22] Sameer Borwankar, Warut Khern-am nuai, and Karthik Natarajan Kannan. 2024. Unraveling the impact: An empirical investigation of ChatGPT's exclusion from stack overflow. Available at SSRN 4481959 (2024).

[23] Rafael Brandão, Joel Carbonera, Clarisse de Souza, Juliana Ferreira, Bernardo Gonçalves, and Carla Leitão. 2019. Mediation Challenges and Socio-Technical Gaps for Explainable Deep Learning Applications. https://arxiv.org/abs/1907.07178 _eprint: 1907.07178.

[24] Tone Bratteteig and Guri Verne. 2018. Does AI make PD obsolete? exploring challenges from artificial intelligence to participatory design. In Proceedings of the 15th Participatory Design Conference: Short Papers, Situated Actions, Workshops and Tutorial - Volume 2 (PDC '18). Association for Computing Machinery, New York, NY, USA. doi:10.1145/3210604.3210646 event-place: Hasselt and Genk, Belgium.

[25] Brocktreee. 2020. [RULE 5 UPDATE] Studies and surveys are no longer allowed. https://www.reddit.com/r/BipolarReddit/comments/f9mp0g/rule_5_update_studies_and_surveys_are_no_longer/

[26] Eliane Bucher, Christian Fieseler, Christoph Lutz, and Alexander Buhmann. 2024. Professionals, purpose-seekers, and passers-through: How microworkers reconcile alienation and platform commitment through identity work. New Media & Society 26, 1 (2024), 190–215. arXiv:https://doi.org/10.1177/14614448211056863 doi:10.1177/14614448211056863

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

[28] Robin Burke, Joseph Konstan, and Michael Ekstrand. 2024. Conducting Recommender Systems User Studies Using POPROX. In Proceedings of the 18th ACM Conference on Recommender Systems (RecSys '24). Association for Computing Machinery, New York, NY, USA, 1277–1278. doi:10.1145/3640457.3687092 event-place: Bari, Italy.

[29] Federico Cabitza, Andrea Campagner, and Clara Balsano. 2020. Bridging the "last mile" gap between AI implementation and operation:"data awareness" that matters. Annals of translational medicine 8, 7 (2020), 501.

[30] Carolin Cardamone, Kevin Schawinski, Marc Sarzi, Steven P. Bamford, Nicola Bennert, C. M. Urry, Chris Lintott, William C. Keel, John Parejko, Robert C. Nichol, Daniel Thomas, Dan Andreescu, Phil Murray, M. Jordan Raddick, Anže Slosar, Alex Szalay, and Jan VandenBerg. 2009. Galaxy Zoo Green Peas: discovery of a class of compact extremely star-forming galaxies*. Monthly Notices of the Royal Astronomical Society 399, 3 (10 2009), 1191–1205. arXiv:https://academic.oup.com/mnras/article-pdf/399/3/1191/18690379/mnras0399-1191.pdf doi:10.1111/j.1365-2966.2009.15383.x

[31] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, and others. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. arXiv preprint arXiv:2307.15217 (2023).

[32] Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. 2025. PERSONA: A Reproducible Testbed for Pluralistic Alignment. In Proceedings of the 31st International Conference on Computational Linguistics, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 11348–11368. https://aclanthology.org/2025.coling-main.752/

[33] Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. 2024. MaxMin-RLHF: Alignment with Diverse Human Preferences. https://arxiv.org/abs/2402.08925 _eprint: 2402.08925.

[34] Stevie Chancellor, Jessica L Feuston, and Jayhyun Chang. 2023. Contextual gaps in machine learning for mental illness prediction: The case of diagnostic disclosures. Proceedings of the ACM on Human-Computer Interaction 7, CSCW2 (2023), 1–27. Publisher: ACM New York, NY, USA.

[35] Shreyas Chaudhari, Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, Ameet Deshpande, and Bruno Castro da Silva. 2025. RLHF Deciphered: A Critical Analysis of Reinforcement Learning from Human Feedback for LLMs. ACM Comput. Surv. 58, 2 (Sept. 2025). doi:10.1145/3743127 Place: New York, NY, USA Publisher: Association for Computing Machinery.

[36] Kai Chen, Zihao He, Taiwei Shi, and Kristina Lerman. 2025. STEER-BENCH: A Benchmark for Evaluating the Steerability of Large Language Models. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 18327–18355. doi:10.18653/v1/2025.emnlp-main.925

[37] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. Advances in neural information processing systems 30 (2017).

[38] Ned Cooper and Alexandra Zafiroglu. 2024. From Fitting Participation to Forging Relationships: The Art of Participatory ML. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24). Association for Computing Machinery, New York, NY, USA. doi:10.1145/3613904.3642775 event-place: Honolulu, HI, USA.

[39] 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 (EAAMO '23). Association for Computing Machinery, New York, NY, USA. doi:10.1145/3617694.3623228 event-place: Boston, MA, USA.

[40] Sasha Costanza-Chock. 2020. Design justice: Community-led practices to build the worlds we need. The MIT Press.

[41] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: user lifecycle and linguistic change in online communities. In Proceedings of the 22nd International Conference on World Wide Web (Rio de Janeiro, Brazil) (WWW '13). Association for Computing Machinery, New York, NY, USA, 307–318. doi:10.1145/2488388.2488416

[42] Christopher A Le Dantec and Carl DiSalvo. 2013. Infrastructuring and the formation of publics in participatory design. Social Studies of Science 43, 2 (2013), 241–264. Publisher: SAGE Publications Sage UK: London, England.

[43] Arun Das and Paul Rad. 2020. Opportunities and challenges in explainable artificial intelligence (xai): A survey. arXiv preprint arXiv:2006.11371 (2020).

[44] 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. 1–23.

[45] Shachar Don-Yehiya, Ben Burtenshaw, Ramon Fernandez Astudillo, Cailean Osborne, Mimansa Jaiswal, Tzu-Sheng Kuo, Wenting Zhao, Idan Shenfeld, Andi Peng, Mikhail Yurochkin, Atoosa Kasirzadeh, Yangsibo Huang, Tatsunori Hashimoto, Yacine Jernite, Daniel Vila-Suero, Omri Abend, Jennifer Ding, Sara Hooker, Hannah Rose Kirk, and Leshem Choshen. 2025. The future of open human feedback. Nature Machine Intelligence 7, 6 (June 2025), 825–835. doi:10.1038/s42256-025-01038-2

[46] Joseph Donia and Jay Shaw. 2021. Co-design and Ethical Artificial Intelligence for Health: Myths and Misconceptions. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (Virtual Event, USA) (AIES '21). Association for Computing Machinery, New York, NY, USA, 77. doi:10.1145/3461702.3462537

[47] Michael Feffer, Nikolas Martelaro, and Hoda Heidari. 2023. The AI Incident Database as an Educational Tool to Raise Awareness of AI Harms: A Classroom Exploration of Efficacy, Limitations, & Future Improvements. In Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23). Association for Computing Machinery, New York, NY, USA. doi:10.1145/3617694.3623223 event-place: Boston, MA, USA.

[48] Michael Feffer, Michael Skirpan, Zachary Lipton, and Hoda Heidari. 2023. From Preference Elicitation to Participatory ML: A Critical Survey & Guidelines for Future Research. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23). Association for Computing Machinery, New York, NY, USA, 38–48. doi:10.1145/3600211.3604661 event-place: Montréal, QC, Canada.

[49] Dror G. Feitelson. 2023. "We do not appreciate being experimented on": Developer and researcher views on the ethics of experiments on open-source projects. Journal of Systems and Software 204 (2023), 111774. doi:10.1016/j.jss.2023.111774

[50] Casey Fiesler and Amy S. Bruckman. 2019. Creativity, Copyright, and Close-Knit Communities: A Case Study of Social Norm Formation and Enforcement. Proc. ACM Hum.-Comput. Interact. 3, GROUP (Dec. 2019). doi:10.1145/3361122

[51] Casey Fiesler, Michael Zimmer, Nicholas Proferes, Sarah Gilbert, and Naiyan Jones. 2024. Remember the Human: A Systematic Review of Ethical Considerations in Reddit Research. Proc. ACM Hum.-Comput. Interact. 8, GROUP (Feb. 2024). doi:10.1145/3633070 Place:

New York, NY, USA Publisher: Association for Computing Machinery.
[52] Heather Ford and Michael Davis. [n. d.]. Implications of generative AI for knowledge integrity on Wikipedia. ([n. d.]).
[53] Deen Freelon. 2015. Discourse architecture, ideology, and democratic norms in online political discussion. New Media & Society 17, 5 (2015), 772–791. arXiv:https://doi.org/10.1177/1461444813513259 doi:10.1177/1461444813513259
[54] Dennis Friess and Christiane Eilders. 2015. A Systematic Review of Online Deliberation Research. Policy & Internet 7, 3 (2015), 319–339. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/poi3.95 doi:10.1002/poi3.95
[55] Chris Fullwood, Darren Chadwick, Melanie Keep, Alison Attrill-Smith, Titus Asbury, and Grainne Kirwan. 2019. Lurking towards empowerment: Explaining propensity to engage with online health support groups and its association with positive outcomes. Computers in Human Behavior 90 (Jan. 2019), 131–140. doi:10.1016/j.chb.2018.08.037
[56] John Gastil and Laura Black. 2007. Public deliberation as the organizing principle of political communication research. Journal of Public Deliberation 4, 1 (2007).
[57] Matthew Groh. 2022. Identifying the Context Shift between Test Benchmarks and Production Data. https://arxiv.org/abs/2207.01059 _eprint: 2207.01059.
[58] hacksoncode. 2025. Research and changemyview. https://www.reddit.com/r/changemyview/wiki/research/
[59] Aaron Halfaker and R Stuart Geiger. 2020. Ores: Lowering barriers with participatory machine learning in wikipedia. Proceedings of the ACM on Human-Computer Interaction 4, CSCW2 (2020), 1–37. Publisher: ACM New York, NY, USA.
[60] 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. doi:10.1145/3706599.3706683
[61] F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. Acm transactions on interactive intelligent systems (tiis) 5, 4 (2015), 1–19. Publisher: Acm New York, NY, USA.
[62] Richard Harper and Dave Randall. 2024. Machine Learning and the Work of the User. Computer Supported Cooperative Work (CSCW) 33, 2 (June 2024), 103–136. doi:10.1007/s10606-023-09483-6
[63] Christina Harrington, Sheena Erete, and Anne Marie Piper. 2019. Deconstructing Community-Based Collaborative Design: Towards More Equitable Participatory Design Engagements. Proc. ACM Hum.-Comput. Interact. 3, CSCW (Nov. 2019). doi:10.1145/3359318 Place: New York, NY, USA Publisher: Association for Computing Machinery.
[64] Thomas T Hills, Peter M Todd, David Lazer, A David Redish, and Iain D Couzin. 2015. Exploration versus exploitation in space, mind, and society. Trends in cognitive sciences 19, 1 (2015), 46–54. Publisher: Elsevier.
[65] Anna Lauren Hoffmann. 2019. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. Information, Communication & Society 22, 7 (2019), 900–915. Publisher: Taylor & Francis.
[66] 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. 1395–1417.
[67] Sohyeon Hwang and Jeremy D. Foote. 2021. Why do People Participate in Small Online Communities? Proc. ACM Hum.-Comput. Interact. 5, CSCW2 (Oct. 2021). doi:10.1145/3479606
[68] Jane Im, Amy X. Zhang, Christopher J. Schilling, and David Karger. 2018. Deliberation and Resolution on Wikipedia: A Case Study of Requests for Comments. Proc. ACM Hum.-Comput. Interact. 2, CSCW, Article 74 (Nov. 2018), 24 pages. doi:10.1145/3274343
[69] Megan N Imundo and David N Rapp. 2022. When fairness is flawed: Effects of false balance reporting and weight-of-evidence statements on beliefs and perceptions of climate change. Journal of Applied Research in Memory and Cognition 11, 2 (2022), 258. Publisher: Educational Publishing Foundation.
[70] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, and others. 2023. Ai alignment: A comprehensive survey. arXiv preprint arXiv:2310.19852 (2023).
[71] Philip N. Johnson-Laird and Marco Ragni. 2023. What Should Replace the Turing Test? Intelligent Computing 2 (2023), 0064. doi:10.34133/icomputing.0064 _eprint: https://spj.science.org/doi/pdf/10.34133/icomputing.0064.
[72] Donald Martin Jr, Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S. Isaac. 2020. Participatory Problem Formulation for Fairer Machine Learning Through Community Based System Dynamics. https://arxiv.org/abs/2005.07572 _eprint: 2005.07572.
[73] Michael Kaschesky and Reinhard Riedl. 2009. Top-level decisions through public deliberation on the internet: evidence from the evolution of Java governance. democracy 4 (2009), 40.
[74] Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2024. A survey of reinforcement learning from human feedback. (2024).
[75] Vijay Keswani, Vincent Conitzer, Hoda Heidari, Jana Schaich Borg, and Walter Sinnott-Armstrong. 2024. On the Pros and Cons of Active Learning for Moral Preference Elicitation. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, Vol. 7. 711–723.

[76] Vijay Keswani, Cyrus Cousins, Breanna Nguyen, Vincent Conitzer, Hoda Heidari, Jana Schaich Borg, and Walter Sinnott-Armstrong. 2025. Moral Change or Noise? On Problems of Aligning AI With Temporally Unstable Human Feedback. https://arxiv.org/abs/2511.10032 _eprint: 2511.10032.

[77] Malik Khadar, Daniel Runningen, Julia Tang, Stevie Chancellor, and Harmanpreet Kaur. 2025. Wisdom of the Crowd, Without the Crowd: A Socratic LLM for Asynchronous Deliberation on Perspectivist Data. Proc. ACM Hum.-Comput. Interact. 9, 7, Article CSCW526 (Oct. 2025), 35 pages. doi:10.1145/3757707

[78] Hannah Rose Kirk, Andrew M Bean, Bertie Vidgen, Paul Röttger, and Scott A Hale. 2023. The past, present and better future of feedback learning in large language models for subjective human preferences and values. arXiv preprint arXiv:2310.07629 (2023).

[79] Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. 2023. The empty signifier problem: Towards clearer paradigms for operationalising" alignment" in large language models. arXiv preprint arXiv:2310.02457 (2023).

[80] Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. 2024. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. Nature Machine Intelligence 6, 4 (2024), 383–392. Publisher: Nature Publishing Group UK London.

[81] Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M Bean, Katerina Margatina, Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, and others. 2024. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. Advances in Neural Information Processing Systems 37 (2024), 105236–105344.

[82] Tianqi Kou, Dana Calacci, and Cindy Lin. 2025. Dead Zone of Accountability: Why Social Claims in Machine Learning Research Should Be Articulated and Defended. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, Vol. 8. 1501–1512. Issue: 2.

[83] Kelsey Kraus and Margaret Kroll. 2025. Maximizing Signal in Human-Model Preference Alignment. Proceedings of the AAAI Conference on Artificial Intelligence 39, 26 (April 2025), 27392–27400. doi:10.1609/aaai.v39i26.34950

[84] Robert E Kraut and Paul Resnick. 2012. Building successful online communities: Evidence-based social design. Mit Press.

[85] Max Krüger, Ana Bustamante Duarte, Anne Weibert, Konstantin Aal, Reem Talhouk, and Oussama Metatla. 2019. What is participation? emerging challenges for participatory design in globalized conditions. Interactions 26, 3 (April 2019), 50–54. doi:10.1145/3319376

[86] 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 (CHI '24). Association for Computing Machinery, New York, NY, USA. doi:10.1145/3613904.3642278 event-place: Honolulu, HI, USA.

[87] Nathan Lambert, Lewis Tunstall, Nazneen Rajani, and Tristan Thrush. 2023. HuggingFace H4 Stack Exchange Preference Dataset. https://huggingface.co/datasets/HuggingFaceH4/stack-exchange-preferences

[88] Cliff Lampe, Rick Wash, Alcides Velasquez, and Elif Ozkaya. 2010. Motivations to participate in online communities. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10). Association for Computing Machinery, New York, NY, USA, 1927–1936. doi:10.1145/1753326.1753616

[89] Hélène Landemore and Scott E. Page. 2015. Deliberation and disagreement: Problem solving, prediction, and positive dissensus. Politics, Philosophy & Economics 14, 3 (2015), 229–254. arXiv:https://doi.org/10.1177/1470594X14544284 doi:10.1177/1470594X14544284

[90] Bruno Latour. 2005. Reassembling the social: An introduction to actor-network-theory. Oxford university press.

[91] 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 (Nov. 2019). doi:10.1145/3359283 Place: New York, NY, USA Publisher: Association for Computing Machinery.

[92] Joran Leest, Claudia Raibulet, Patricia Lago, and Ilias Gerostathopoulos. 2025. From Tea Leaves to System Maps: A Survey and Framework on Context-aware Machine Learning Monitoring. IEEE Transactions on Software Engineering (2025), 1–30. doi:10.1109/TSE.2025.3602520

[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] Changye Li, Zachary Levonian, Haiwei Ma, and Svetlana Yarosh. 2018. Condition Unknown: Predicting Patients' Health Conditions in an Online Health Community. In Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '18 Companion). Association for Computing Machinery, New York, NY, USA, 281–284. doi:10.1145/3272973.3274075 event-place: Jersey City, NJ, USA.

[95] Chu Li, Rock Yuren Pang, Delphine Labbé, Yochai Eisenberg, Maryam Hosseini, and Jon E. Froehlich. 2025. Accessibility for Whom? Perceptions of Mobility Barriers Across Disability Groups and Implications for Designing Personalized Maps. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25). Association for Computing Machinery, New York, NY, USA, Article 44, 19 pages. doi:10.1145/3706598.3713421

[96] 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. doi:10.1145/3706598.3713292

[97] Jonathan K. London, Krista A. Haapanen, Ann Backus, Savannah M. Mack, Marti Lindsey, and Karen Andrade. 2020. Aligning Community-Engaged Research to Context. International Journal of Environmental Research and Public Health 17, 4 (2020). doi:10.3390/ijerph17041187

[98] Lena Mamykina, Drashko Nakikj, and Noemie Elhadad. 2015. Collective Sensemaking in Online Health Forums. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15). Association for Computing Machinery, New York, NY, USA, 3217–3226. doi:10.1145/2702123.2702566

[99] Noortje Marres, Michael Castelle, Beatrice Gobbo, Chiara Poletti, and James Tripp. 2024. AI as super-controversy: Eliciting AI and society controversies with an extended expert community in the UK. Big Data & Society 11, 2 (2024), 20539517241255103. arXiv:https://doi.org/10.1177/20539517241255103 doi:10.1177/20539517241255103

[100] J. Nathan Matias and Merry Mou. 2018. CivilServant: Community-Led Experiments in Platform Governance. 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, 1–13. doi:10.1145/3173574.3173583

[101] J. Nathan Matias and Megan Price. 2025. How public involvement can improve the science of AI. Proceedings of the National Academy of Sciences 122, 48 (Dec. 2025), e2421111122. doi:10.1073/pnas.2421111122 Publisher: Proceedings of the National Academy of Sciences.

[102] Melissa McCradden, Katrina Hui, and Daniel Z Buchman. 2023. Evidence, ethics and the promise of artificial intelligence in psychiatry. Journal of medical ethics 49, 8 (2023), 573–579. Publisher: Institute of Medical Ethics.

[103] Sean McGregor. 2021. Preventing repeated real world AI failures by cataloging incidents: The AI incident database. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35. 15458–15463. Issue: 17.

[104] Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. Proc. ACM Hum.-Comput. Interact. 4, CSCW2 (Oct. 2020). doi:10.1145/3415186 Place: New York, NY, USA Publisher: Association for Computing Machinery.

[105] Chris J Michael, Dina Acklin, and Jaelle Scheuerman. 2020. On interactive machine learning and the potential of cognitive feedback. arXiv preprint arXiv:2003.10365 (2020).

[106] Dang Minh, H. Xiang Wang, Y. Fen Li, and Tan N. Nguyen. 2022. Explainable artificial intelligence: a comprehensive review. Artificial Intelligence Review 55, 5 (June 2022), 3503–3568. doi:10.1007/s10462-021-10088-y

[107] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2023. Human-in-the-loop machine learning: a state of the art. Artificial Intelligence Review 56, 4 (April 2023), 3005–3054. doi:10.1007/s10462-022-10246-w

[108] Michael Muller and Allison Druin. 2002. Participatory Design: The Third Space in HCI. Handbook of HCI (Jan. 2002).

[109] Dylan Mulvin. 2021. Proxies: The Cultural Work of Standing In. The MIT Press. doi:10.7551/mitpress/11765.001.0001

[110] Yiftach Nagar. 2012. What do you think? the structuring of an online community as a collective-sensemaking process. In Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12). Association for Computing Machinery, New York, NY, USA, 393–402. doi:10.1145/2145204.2145266

[111] Yu-Leung Ng. 2024. A longitudinal model of continued acceptance of conversational artificial intelligence. Information Technology & People 38, 4 (April 2024), 1871–1889. doi:10.1108/ITP-06-2023-0577

[112] Ritesh Noothigattu, Snehalkumar 'Neil' S. Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel D. Procaccia. 2018. A Voting-Based System for Ethical Decision Making. https://arxiv.org/abs/1709.06692 _eprint: 1709.06692.

[113] Alexandra Olteanu, Su Lin Blodgett, Agathe Balayn, Angelina Wang, Fernando Diaz, Flavio du Pin Calmon, Margaret Mitchell, Michael Ekstrand, Reuben Binns, and Solon Barocas. 2025. Rigor in AI: Doing Rigorous AI Work Requires a Broader, Responsible AI-Informed Conception of Rigor. arXiv preprint arXiv:2506.14652 (2025).

[114] Edeh Michael Onyema, Khalid K Almuzaini, Fergus Uchenna Onu, Devvret Verma, Ugboaja Samuel Gregory, Monika Puttaramaiah, and Rockson Kwasi Afriyie. 2022. Prospects and challenges of using machine learning for academic forecasting. Computational Intelligence and Neuroscience 2022, 1 (2022), 5624475. Publisher: Wiley Online Library.

[115] Katherine Panciera, Aaron Halfaker, and Loren Terveen. 2009. Wikipedians are born, not made: a study of power editors on Wikipedia. In Proceedings of the 2009 ACM international conference on supporting group work. 51–60.

[116] Zizi Papacharissi. 2004. Democracy online: civility, politeness, and the democratic potential of online political discussion groups. New Media & Society 6, 2 (2004), 259–283. arXiv:https://doi.org/10.1177/1461444804041444 doi:10.1177/1461444804041444

[117] phareous. 2023. Rule Change: No More Surveys. https://www.reddit.com/r/adhd_anxiety/comments/13vueds/rule_change_no_more_surveys

[118] Lisa Posch, Arnim Bleier, Clemens M. Lechner, Daniel Danner, Fabian Flöck, and Markus Strohmaier. 2019. Measuring Motivations of Crowdworkers: The Multidimensional Crowdworker Motivation Scale. Trans. Soc. Comput. 2, 2, Article 8 (Sept. 2019), 34 pages. doi:10.1145/3335081

[119] Vinodkumar Prabhakaran and Donald Jr Martin. 2020. Participatory Machine Learning Using Community-Based System Dynamics. Health and human rights 22, 2 (Dec. 2020), 71–74. Place: United States.

[120] Gale H. Prinster, C. Estelle Smith, Chenhao Tan, and Brian C. Keegan. 2024. Community Archetypes: An Empirical Framework for Guiding Research Methodologies to Reflect User Experiences of Sense of Virtual Community on Reddit. Proc. ACM Hum.-Comput. Interact. 8, CSCW1 (April 2024). doi:10.1145/3637310 Place: New York, NY, USA Publisher: Association for Computing Machinery.

[121] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. https://arxiv.org/abs/2305.18290 _eprint: 2305.18290.

[122] Gonzalo Ramos, Christopher Meek, Patrice Simard, Jina Suh, and Soroush Ghorashi. 2020. Interactive machine teaching: a human-centered approach to building machine-learned models. Human–Computer Interaction 35, 5-6 (Nov. 2020), 413–451. doi:10.1080/07370024.2020.1734931 Publisher: Taylor & Francis.

[123] Samantha Robertson and Niloufar Salehi. 2020. What If I Don't Like Any Of The Choices? The Limits of Preference Elicitation for Participatory Algorithm Design. https://arxiv.org/abs/2007.06718 _eprint: 2007.06718.

[124] Samantha Robertson and Niloufar Salehi. 2020. What if I don't like any of the choices? The limits of preference elicitation for participatory algorithm design. arXiv preprint arXiv:2007.06718 (2020).

[125] David Rolnick, Alan Aspuru-Guzik, Sara Beery, Bistra Dilkina, Priya L. Donti, Marzyeh Ghassemi, Hannah Kerner, Claire Monteleoni, Esther Rolf, Milind Tambe, and Adam White. 2024. Position: Application-Driven Innovation in Machine Learning. In Forty-first International Conference on Machine Learning. https://openreview.net/forum?id=xEB2oF3vvb

[126] Sabirat Rubya, Joseph Numainville, and Svetlana Yarosh. 2021. Comparing Generic and Community-Situated Crowdsourcing for Data Validation in the Context of Recovery from Substance Use Disorders. 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 449, 17 pages. doi:10.1145/3411764.3445399

[127] Sabirat Rubya, Joseph Numainville, and Svetlana Yarosh. 2021. Comparing Generic and Community-Situated Crowdsourcing for Data Validation in the Context of Recovery from Substance Use Disorders. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA. doi:10.1145/3411764.3445399 event-place: Yokohama, Japan.

[128] Abdullah Hasan Safir, Noshin Tahsin, Pratyasha Saha, Dipannita Nandi, Zulkarin Jahangir, Cecily Morrison, Syed Ishtiaque Ahmed, and Nusrat Jahan Mim. 2025. Collective Agency in Art-making: Towards Community-centric Design of Text-to-Image (T2I) AI Tools. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 8, 3 (Oct. 2025), 2248–2260. doi:10.1609/aies.v8i3.36710

[129] Bukhoree Sahoh and Anant Choksuriwong. 2023. The role of explainable Artificial Intelligence in high-stakes decision-making systems: a systematic review. Journal of ambient intelligence and humanized computing 14, 6 (2023), 7827–7843. doi:10.1007/s12652-023-04594-w Place: Germany.

[130] Subramanyam Sahoo, Aman Chadha, Vinija Jain, and Divya Chaudhary. 2025. Position: The Complexity of Perfect AI Alignment – Formalizing the RLHF Trilemma. https://arxiv.org/abs/2511.19504 _eprint: 2511.19504.

[131] Lynn M Sanders. 1997. Against deliberation. Political theory 25, 3 (1997), 347–376.

[132] Gilbert Sarah. 2025. Guidelines for Research. https://www.reddit.com/r/AskHistorians/wiki/research/

[133] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. Irresolvable Disagreement: A Study on Worker Deliberation in Crowd Work. Proc. ACM Hum.-Comput. Interact. 2, CSCW, Article 154 (Nov. 2018), 19 pages. doi:10.1145/3274423

[134] Doug Schuler. 1994. Community networks: building a new participatory medium. Commun. ACM 37, 1 (Jan. 1994), 38–51. doi:10.1145/175222.175225

[135] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Association for Computing Machinery, New York, NY, USA, 59–68. doi:10.1145/3287560.3287598

[136] Shreya Shankar, Rolando Garcia, Joseph M. Hellerstein, and Aditya G. Parameswaran. 2024. "We Have No Idea How Models will Behave in Production until Production": How Engineers Operationalize Machine Learning. Proc. ACM Hum.-Comput. Interact. 8, CSCW1 (April 2024). doi:10.1145/3653697 Place: New York, NY, USA Publisher: Association for Computing Machinery.

[137] Shreya Shankar, J.D. Zamfirescu-Pereira, Bjoern Hartmann, Aditya Parameswaran, and Ian Arawjo. 2024. Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences. In Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24). Association for Computing Machinery, New York, NY, USA. doi:10.1145/3654777.3676450 event-place: Pittsburgh, PA, USA.

[138] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, and others. 2023. Towards understanding sycophancy in language models. arXiv preprint arXiv:2310.13548 (2023).

[139] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and others. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. 723–741.

[140] Murtuza N Shergadwala, Himabindu Lakkaraju, and Krishnaram Kenthapadi. 2022. A Human-Centric Perspective on Model Monitoring. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 10, 1 (Oct. 2022), 173–183. doi:10.1609/hcomp.v10i1.21997

[141] Zheyuan Ryan Shi, Leah Lizarondo, and Fei Fang. 2021. A Recommender System for Crowdsourcing Food Rescue Platforms. In Proceedings of the Web Conference 2021 (WWW '21). Association for Computing Machinery, New York, NY, USA, 857–865. doi:10.1145/3442381.3449787 event-place: Ljubljana, Slovenia.

[142] Ruth Shortall, Anatol Itten, Michiel van der Meer, Pradeep Murukannaiah, and Catholijn Jonker. 2022. Reason against the machine? Future directions for mass online deliberation. Frontiers in Political Science 4 (2022), 946589. Publisher: Frontiers Media SA.

[143] Jan Simson, Fiona Draxler, Samuel Mehr, and Christoph Kern. 2025. Preventing harmful data practices by using participatory input to navigate the machine learning multiverse. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. 1–30.

[144] 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.

[145] Paul Slovic. 1995. The construction of preference. American psychologist 50, 5 (1995), 364. Publisher: American Psychological Association.

[146] Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2024. Preference ranking optimization for human alignment. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38. 18990–18998. Issue: 17.

[147] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, and others. 2024. A roadmap to pluralistic alignment. arXiv preprint arXiv:2402.05070 (2024).

[148] Karolina Stańczak, Nicholas Meade, Mehar Bhatia, Hattie Zhou, Konstantin Böttinger, Jeremy Barnes, Jason Stanley, Jessica Montgomery, Richard Zemel, Nicolas Papernot, Nicolas Chapados, Denis Therien, Timothy P. Lillicrap, Ana Marasović, Sylvie Delacroix, Gillian K. Hadfield, and Siva Reddy. 2025. Societal Alignment Frameworks Can Improve LLM Alignment. https://arxiv.org/abs/2503.00069 _eprint: 2503.00069.

[149] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 3008–3021. https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf

[150] Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21). Association for Computing Machinery, New York, NY, USA. doi:10.1145/3465416.3483305 event-place: –, NY, USA.

[151] Harini Suresh, Rajiv Movva, Amelia Lee Dogan, Rahul Bhargava, Isadora Cruxen, Angeles Martinez Cuba, Guilia Taurino, Wonyoung So, and Catherine D'Ignazio. 2022. Towards intersectional feminist and participatory ML: A case study in supporting feminicide counterdata collection. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 667–678.

[152] 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 (FAccT '24). Association for Computing Machinery, New York, NY, USA, 1609–1621. doi:10.1145/3630106.3658992 event-place: Rio de Janeiro, Brazil.

[153] Bárbara Szaniecki, Bibiana Serpa, Imaíra Portela, Marina Sirito, Mariana Costard, and Sâmia Batista. 2020. Participation otherwise: Practices by/from the Global South. In Proceedings of the 16th Participatory Design Conference 2020 - Participation(s) Otherwise - Volume 2 (Manizales, Colombia) (PDC '20). Association for Computing Machinery, New York, NY, USA, 203–205. doi:10.1145/3384772.3385171

[154] Mohammad Tahaei, Daricia Wilkinson, Alisa Frik, Michael Muller, Ruba Abu-Salma, and Lauren Wilcox. 2024. Surveys considered harmful? Reflecting on the use of surveys in AI research, development, and governance. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, Vol. 7. 1416–1433.

[155] Dario Taraborelli and Giovanni Luca Ciampaglia. 2010. Beyond Notability. Collective Deliberation on Content Inclusion in Wikipedia. In 2010 Fourth IEEE International Conference on Self-Adaptive and Self-Organizing Systems Workshop. 122–125. doi:10.1109/SASOW.2010.26

[156] Niels Christian Taubert. 2008. Balancing requirements of decision and action: Decision-making and implementation in free/open source software projects. Science, Technology & Innovation Studies 4, 1 (2008).

[157] Tengrrl. 2025. CFP: In defense of the commons. https://wpa-announcements.tracigardner.com/2025/12/04/cfp-in-defense-of-the-commons/ Publication Title: WPA.

[158] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. https://arxiv.org/abs/2302.13971 _eprint: 2302.13971.

[159] Thi Phuong Thao Tran, Jacque-Corey Cormier, Corey Anthony Hopwood, Jordan Foster, Isabel Scheib, Fei Li, Kathleen A Dolan, Nicole A Lynch, Dawn M Aycock, Claire A Spears, and others. 2025. Building Trust for Community-Engaged Research: Recommendations From

a Qualitative Study. Journal of Participatory Research Methods 6, 2 (2025), 74–98. Publisher: Specialty Publications.

[160] Tommaso Turchi, Daria Mikhaylova, Miriana Troccoli, Alessio Malizia, Mario Giovanni C.A. Cimino, Federico Andrea Galatolo, Gaetano La Mantia, Giuseppina Campisi, and Olga Di Fede. 2025. Ecological Validity Missing in AI-Assisted Clinical Decision Support Research: Why Real-World Context Matters. In Proceedings of the 16th Biannual Conference of the Italian SIGCHI Chapter (CHItaly '25). Association for Computing Machinery, New York, NY, USA. doi:10.1145/3750069.3750072

[161] Gael Varoquaux, Sasha Luccioni, and Meredith Whittaker. 2025. Hype, Sustainability, and the Price of the Bigger-is-Better Paradigm in AI. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25). Association for Computing Machinery, New York, NY, USA, 61–75. doi:10.1145/3715275.3732006

[162] Patricia Vázquez-Villegas, María Del Pilar García-Chitiva, Danilo Valdes-Ramirez, Carmen Isabel Reyes Peraza, Carles Abarca de Haro, and Genaro Zavala. 2024. WIP: Implementing and Deploying Artificial Intelligence Solutions in Higher Education Institutions. In 2024 IEEE Frontiers in Education Conference (FIE). 1–5. doi:10.1109/FIE61694.2024.10893121

[163] Caleb Warren, A Peter McGraw, and Leaf Van Boven. 2011. Values and preferences: Defining preference construction. Wiley Interdisciplinary Reviews: Cognitive Science 2, 2 (2011), 193–205.

[164] Caleb Warren, A Peter McGraw, and Leaf Van Boven. 2011. Values and preferences: defining preference construction. Wiley Interdisciplinary Reviews: Cognitive Science 2, 2 (2011), 193–205. Publisher: Wiley Online Library.

[165] Daniel Weinstock. 2017. Compromise, pluralism, and deliberation. Critical Review of International Social and Political Philosophy 20, 5 (Sept. 2017), 636–655. doi:10.1080/13698230.2017.1328093 Publisher: Routledge.

[166] Galen Weld, Amy X Zhang, and Tim Althoff. 2024. Making online communities 'better': a taxonomy of community values on reddit. In Proceedings of the International AAAI Conference on Web and Social Media, Vol. 18. 1611–1633.

[167] Marty J Wolf, Keith Miller, and Frances S Grodzinsky. 2017. Why we should have seen that coming: comments on Microsoft's tay" experiment," and wider implications. Acm Sigcas Computers and Society 47, 3 (2017), 54–64. Publisher: ACM New York, NY, USA.

[168] Kate Wong. 2025. Google AI grant to inaturalist prompts community outcry. https://www.scientificamerican.com/article/google-ai-grant-to-inaturalist-prompts-community-outcry/ Publication Title: Scientific American.

[169] Scott Wright and John Street. 2007. Democracy, deliberation and design: the case of online discussion forums. New Media & Society 9, 5 (2007), 849–869. arXiv:https://doi.org/10.1177/1461444807081230 doi:10.1177/1461444807081230

[170] Stephen Tze-Inn Wu, Daniel Demetriou, and Rudwan Ali Husain. 2023. Honor Ethics: The Challenge of Globalizing Value Alignment in 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, 593–602. doi:10.1145/3593013.3594026

[171] Chloe Xiang. 2023. Startup Uses AI Chatbot to Provide Mental Health Counseling and Then Realizes It 'Feels Weird'. https://www.vice.com/en/article/startup-uses-ai-chatbot-to-provide-mental-health-counseling-and-then-realizes-it-feels-weird/ Publication Title: Vice.

[172] Matthew Zent, Seraphina Yong, Dhruv Bala, Stevie Chancellor, Joseph A Konstan, Loren Terveen, and Svetlana Yarosh. 2025. Beyond the Individual: A Community-Engaged Framework for Ethical Online Community Research. Proceedings of the ACM on Human-Computer Interaction 9, 7 (2025), 1–33. Publisher: ACM New York, NY, USA.

[173] Amy X Zhang, Michael Muller, and Dakuo Wang. 2020. How do data science workers collaborate? roles, workflows, and tools. Proceedings of the ACM on Human-Computer Interaction 4, CSCW1 (2020), 1–23.

[174] Zhehao Zhang, Ryan A Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, and others. 2024. Personalization of large language models: A survey. arXiv preprint arXiv:2411.00027 (2024).

[175] Kaitlyn Zhou, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman, and Alexandra Olteanu. 2022. Deconstructing NLG Evaluation: Evaluation Practices, Assumptions, and Their Implications. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 314–324. doi:10.18653/v1/2022.naacl-main.24

[176] 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