Decentralising LLM Alignment: A Case for Context, Pluralism, and Participation

Oriane Peter, Kate Devlin

Department of Digital Humanities, King's College London, United Kingdom
London WC2R 2LS, UK
oriane.peter@kcl.ac.uk

Abstract

Large Language Models (LLMs) alignment methods have been credited with the commercial success of products like ChatGPT, given their role in steering LLMs towards user-friendly outputs. However, current alignment techniques predominantly mirror the normative preferences of a narrow reference group, effectively imposing their values on a wide user base. Drawing on theories of the power/knowledge nexus, this work argues that current alignment practices centralise control over knowledge production and governance within already influential institutions. To counter this, we propose decentralising alignment through three characteristics: context, pluralism, and participation. Furthermore, this paper demonstrates the critical importance of delineating the context-of-use when shaping alignment practices by grounding each of these features in concrete use cases. This work makes the following contributions: (1) highlighting the role of context, pluralism, and participation in decentralising alignment; (2) providing concrete examples to illustrate these strategies; and (3) demonstrating the nuanced requirements associated with applying alignment across different contexts of use. Ultimately, this paper positions LLM alignment as a potential site of resistance against epistemic injustice and the erosion of democratic processes, while acknowledging that these strategies alone cannot substitute for broader societal changes.

Introduction

Self-reportedly, the *"secret sauce"* behind OpenAI's commercial success was their introduction of a novel alignment algorithm for Large Language Models (LLMs) (Ouyang et al. 2022; Heaven 2023). This algorithm enables developers to steer LLMs' outputs away from mere pre-trained text completion and towards a more conversational, user-friendly style. It allows developers to avoid, to some extent, the pitfalls faced by early LLMs, such as the replication of 'toxic' views present in the model training data (e.g. Victor 2016). The success of OpenAI's Reinforcement Learning from Human Feedback (RLHF) method has catalysed the development of the LLM alignment research field. However, despite its increasing prevalence in research, the term 'alignment' lacks a consistent definition, leading to critiques that it serves as a rhetorical placeholder more than as a precise concept (Kirk et al. 2023). Such vagueness leaves significant ambiguity about *what* models should be aligned *to* and, critically, *with whom* they should be aligned (Gabriel 2020).

Despite their intended global and diverse applications, LLMs predominantly generate outputs that are reflective of a narrow societal segment: Western, Educated, Industrialised, Rich, and Democratic (WEIRD) societies (Atari et al. 2023; Henrich, Heine, and Norenzayan 2010). While the training data contributes to this bias, the alignment process significantly amplifies this skew. RLHF demonstrably reduces output diversity by collapsing the probability distributions of words around the alignment data. This 'mode collapse' restricts the pluralism of possible model outputs, confining them to the preferences of the target group used for alignment (Casper et al. 2023; Kirk et al. 2024c; Sorensen et al. 2024; O'Mahony et al. 2024). This fundamental constraint of alignment is recognised in OpenAI's initial RLHF's paper, which states that *"it is impossible that one can train a system that is aligned to everyone's preferences at once, or where everyone would endorse the tradeoffs"* (Ouyang et al. 2022, p. 18). Consequently, Ouyang et al. (2022) purposefully confined their model's alignment to *"a specific human reference group for a specific application"*, admitting that their model embodies their own preferences as well as those of their labellers. Likewise, Anthropic acknowledges developers' *"outsized role"* in establishing their models' alignment principles (Anthropic 2023). Ultimately, alignment with human preferences, or the values that underpin them, cannot be universal or generic. Cultures and contexts shape these preferences and values, making it impossible to abstract them entirely. Consequently, any approach to alignment necessarily relies on narrow targets. For communities outside these targets, this leads to forms of *'representational harm'*, including demeaning (e.g., an LLM using derogatory terms targeting marginalised groups), reification (e.g., an LLM reducing complex non-WEIRD identities to rigid, one-dimensional stereotypes), or erasure (e.g., an LLM's outputs consistently omitting certain perspectives from historical narratives) (Gillespie 2024; Blodgett et al. 2020; Katzman et al. 2023).

Efforts to mitigate the skewness of LLM alignment and the harms it generates have frequently focused on developing more inclusive datasets through collaborations with under-represented communities (Rose, Kroner Dale, and

Jusko 2025; Bergman et al. 2024; Kirk et al. 2024b; Huang et al. 2024). However, such inclusion initiatives may not be sufficient to bring about significant changes in power dynamics or meaningfully empower the targeted groups, as we will further discuss in this work (Varshney 2024; Young et al. 2024).

'Representational harms' are not unique to LLMs but rather reflect the systemic inequities and power imbalances embedded in the societies that develop these systems (Suchman and Suchman 2007; Gillespie 2024; Blodgett et al. 2020). Importantly, LLMs do not merely enhance these harms because they are *'misaligned'*; instead, they reproduce them *because of who they are aligned with*. As we discuss in this work, alignment can be viewed as a technological component of a power/knowledge nexus, where institutions leverage their pre-existing power to align LLMs with themselves, leading to LLMs producing forms of knowledge which normalise, stabilise, and ultimately centralise their power. Consequently, LLMs' enhancement and perpetuation of harms result from the enhancement and perpetuation of these power structures (Ganesh and Moss 2022; Varshney 2024; Muldoon and Wu 2023; Mohamed, Png, and Isaac 2020).

Therefore, we propose that meaningfully challenging the perpetuation of harms caused by LLMs necessitates a decentralisation of alignment processes. This decentralisation serves as a means to redistribute the power to control and disseminate information to specific communities, tailored to their use cases. While alignment cannot – and should not – be the *only* site of resistance to hegemony, it remains a valuable point of intervention, provided it incorporates three key characteristics essential for decentralisation: contextualization, pluralism, and participation. Furthermore, the effective implementation of alignment strategies must be grounded in the specific contexts of their deployment. To illustrate this context-dependent nature, we introduce two distinct use cases – a conversational Voter Advice Application (VAA) and an LLM-powered Non-Playable Character (NPC) – chosen for their stark contrast. We then examine how each characteristic may be implemented within these contexts, highlighting their nuances and the necessity for context-specific adaptations.

This paper thus aims to advance discussions on LLM alignment and the power/knowledge nexus in three key ways:

- Emphasising that context, pluralism, and participation are essential for reorienting alignment towards power redistribution.
- Grounding each of these characteristics in a concrete use case.
- Demonstrating the importance of use-case specificity by contrasting the nuanced requirements and implications of different alignment strategies.

The Case for Contextual, Pluralistic, and Participatory Alignment

Our central thesis is that a meaningful decentralisation of power in the context of LLM alignment hinges on three essential characteristics: Context, Pluralism, and Participation. This section lays the foundations of the argument by first presenting the relationship between LLM alignment and Foucault's power/knowledge nexus, and then explaining why each proposed characteristic is crucial for countering centralising tendencies. Next, we trace how these characteristics have been implemented in prior works. Finally, we examine the limitations of these earlier approaches.

LLM Alignment and Power Centralisation

The relationship between knowledge and power has been extensively explored. Our intention, therefore, is not to provide a comprehensive overview of related theories, but rather to present the notions necessary for building our argument. According to Foucault (1975), power and knowledge are mutually constitutive: existing power structures enable the production of certain forms of knowledge, which, in turn, normalise and stabilise those power structures, creating a self-reinforcing cycle. Furthermore, Jasanoff (2006) highlights the deep entanglement of political power with science and technology. She contends that scientific and technological advancements are *"indispensable to the expression and exercise of power"* (Jasanoff 2006, p. 14). This power is concentrated in centres that control knowledge dissemination tools, such as the printing press, enabling specific worldviews to be packaged into *"conveniently portable representations"* (Jasanoff 2006; Latour 1988, p. 22). Consequently, certain forms of hegemonic power can concentrate within institutions that control communication technologies.

To examine the power dynamics they perpetuate, LLMs have been similarly framed as instruments of dissemination and normalisation of hegemonic worldviews. (Durmus et al. 2023; Weidinger et al. 2021, 2022; Cave and Dihal 2020; Bommasani et al. 2022; Gabriel 2020; Jakesch et al. 2023). Given the significant resources and expertise required, training powerful LLMs is typically accessible only to those with pre-existing wealth and influence. These actors can use their resources to shape machine-learning products in ways that promote particular worldviews. In turn, these products help to normalise and spread those perspectives on a global platform. Moreover, LLMs can reproduce, propagate, and normalise specific perspectives at a historically high speed. Their outputs are becoming increasingly indistinguishable from human-generated content and can be easily generated at a massive scale. The extent of this replication is so vast that it raises concerns about the 'pollution' of the online landscape spreading at an unknown scale (Kniaz 2023). Consequently, LLMs not only depend on existing power structures but also centralise hegemonic power in a few institutions.

This centralisation hinges on the ability of companies deploying LLMs to control the outputs of their models. Indeed, these models are used in a variety of subsequent AI products, developed by other companies. They are also trained on extensive online data, which reflects many, though by no means all, human perspectives. Therefore, the control of LLMs as a tool of knowledge dissemination relies on the capability to control *which types of knowledges are disseminated*. This is where alignment comes into play. Alignment

lets developers decide which types of online knowledge a model reproduces and, importantly, which views are suppressed – a moderation process deemed necessary to prevent the replication of toxic or harmful content. However, the determination of what constitutes 'harmful' and 'helpful' rests with these companies. Thus, alignment significantly contributes to the control which powerful institutions hold over LLM-mediated knowledge reproduction, empowering them to promote their own perspectives and consequently normalise and centralise their power (Varshney 2024; Muldoon and Wu 2023; Ganesh and Moss 2022; Mohamed, Png, and Isaac 2020).

The Role of Context, Pluralism, and Participation in Decentralising Power

Context There is a prevalent tendency in machine learning to strive for *'general-purpose systems'* or the creation of *'artificial general intelligence'* (Gebru and Torres 2024; Raji et al. 2021). This tendency is also prevalent in much of the current alignment literature, which often over-generalises the concept of 'humans' (Varshney 2024; Gabriel 2020). Such generality conceals the perspectives and positionalities embedded within LLMs behind a veneer of universality. Moreover, the intention of 'universality' makes these models inherently unsafe, since standard operating conditions, and consequently, measurable harms, cannot be defined (Gebru and Torres 2024). Positionality is also abstracted away from data annotation practices, in which labellers are often framed as interchangeable. Such framing conceals how labellers' lived experiences and often poor working conditions (Miceli and Posada 2022) shape their output, especially when annotating highly subjective topics like harms or values. Therefore, reporting annotation workers' context and accounting for their positionality is essential when generating datasets for applications like alignment (Diaz et al. 2022). Furthermore, context is fundamental to both recognising and intervening in the shift of power dynamics mediated by alignment. It is necessary to explicitly trace *who* is being aligned to, and *for what purpose*. Context helps illuminate the model's impact by revealing what it enables, what it limits, and ultimately, who benefits from its use. Moreover, since alignment practices are inherently tailored to specific groups and applications (Ouyang et al. 2022), context allows for the clear identification of suitable alignment targets within a given case. The availability of open-weight models partly enables this contextualisation, making it possible to tailor AI to specific use cases and communities, from the medical sector (Chen et al. 2023) to meme interpretation (Jha et al. 2024). Therefore, context is necessary to recognise power imbalances, define effective strategies for using alignment to counteract them, and evaluate the success of these efforts.

Pluralism Pluralism, as a political philosophy, advocates for the peaceful coexistence of diverse perspectives. Although not inherently incompatible with alignment, pluralism stands in contrast to the universalist and consensus-driven nature that some current approaches strive for (Varshney 2024). As a result, algorithms like RLHF struggle with managing contradictory goals and even reduce pluralism compared to non-aligned models (Sorensen et al. 2024). Nonetheless, a lack of consensus should not be regarded as a failure that needs to be corrected. On the contrary, Mittelstadt (2019, p. 505) highlights the importance of pluralism in AI ethics by stating:

> *"Ethics is not meant to be easy or formulaic. Intractable principled disagreements should be expected and welcomed, as they reflect both serious ethical consideration and diversity of thought"*.

Thus, aligning LLMs to be contextually relevant to stakeholder cultures and values inherently requires enabling pluralism. Cultures are not fixed, objective entities but rather emerge through ongoing debates, negotiations, and contestations within communities (Qadri et al. 2025). The challenge of navigating this pluralism is not unique to LLM alignment. Online platforms like Reddit or Stack Overflow have similarly grappled with conflicting viewpoints within their communities during content moderation (Mamykina et al. 2011; Raji et al. 2021). Drawing on the mechanisms adopted by these platforms, Kuo et al. (2024) proposes WikiBench, a community-driven curation of Wikipedia data designed to capture the inherent ambiguity and uncertainty beneath the apparent consensus of aggregated community labels. Similar approaches can be incorporated to enable pluralistic alignment, allowing for the coexistence of conflicting values within a single system, promoting diversity of thought, and leaving space for uncertainty. These factors are crucial for preventing the centralisation and homogenization of viewpoints in LLMs (Varshney 2024). Lorde (2007, p. 111) emphasises the necessity of descending *"into the chaos of knowledge"* to *"return with true visions of our future, along with the concomitant power to effect those changes which can bring that future into being"*. This is only possible, she notes, through *"the interdependence of mutual (nondominant) differences"*. Thus, the 'chaos' of dissonant, uncertain and conflicting views should not be smoothed away by alignment, but rather preserved and leveraged for its ability to bridge partial understanding of realities and bring about change.

Participation Participation is often conceptualised as a practice aimed at rebalancing power dynamics between designers and stakeholders, acknowledging the situated nature of knowledge and respecting stakeholders' experiential expertise (Birhane et al. 2022; Young et al. 2024). The growing recognition of the need for participation in AI development marks what Delgado et al. (2023) term the 'participatory turn' in AI design. If alignment is to redistribute power effectively, participatory practices are indispensable, providing stakeholders with agency in shaping the model training, alignment and evaluation. However, participation manifests in diverse forms, not all of which genuinely empower participants. Sloane et al. (2022) cautions against 'participation-washing', the creation of a facade of inclusivity that masks the exploitation of cheap or free annotation labour while denying communities meaningful influence over system design. Similarly, Delgado et al. (2023, p. 4) delineates modes of participation along an axis of involve-

ment, ranging from *"transactional preference elicitation"* to *"transformative subversion of power dynamics"*. However, they avoid a normative ranking of these modes, emphasising that complete ownership transfer should not necessarily be the ultimate goal of participation. The power dynamics of participation are particularly crucial to consider when participation is operationalised as annotation work. This work is often outsourced to marginalised groups, in reportedly poor working conditions, raising serious concerns about the invisibilisation and exploitation of their labour. The pressure of unrealistic workloads combined with strict quality requirements can lead to unpaid overtime when these requirements aren't met. Consequently, annotators tend to prioritise what they perceive as the norms and expectations of the requesters over their own perspectives, highlighting the limitations of hiring diverse annotators as a means to diversify represented viewpoints in a dataset.(Gray and Suri 2019; Wang, Prabhat, and Sambasivan 2022; Le Ludec, Cornet, and Casilli 2023; Muldoon et al. 2023; Miceli, Schuessler, and Yang 2020; Miceli and Posada 2022) . Therefore, Diaz et al. (2022) recommend carefully considering and documenting conditions of work when aggregating annotation data from crowdworkers. Furthermore, participation as a recognition of localised and lived expertise exists in tension with the aforementioned universalist aims of many LLM systems. Therefore, Suresh et al. (2024) emphasises that participation must be grounded in specific domains to mitigate the limitations and frustrations which generalist ideation can entail.

Context, Pluralism, and Participation in Previous Work

Recent efforts to counter the narrow coverage of alignment have focused on creating inclusive datasets for both top–down constitutional AI alignment (e.g., Bai et al. 2022) and bottom–up, RLHF-based approaches (e.g., Ouyang et al. 2022). For example, Huang et al. (2024) have sought to extend Anthropic's AI constitution by incorporating feedback from approximately 1,000 Americans, while Kirk et al. (2024b) developed the PRISM dataset through the collection of AI output preferences from 1,500 English speakers. Similarly, Rose, Kroner Dale, and Jusko (2025) are assembling an open–source RLHF dataset by capturing views on 10 specific use–cases from around 15,000 respondents across 5 countries. In contrast, Bergman et al. (2024) adopted a community–centred approach aimed at norm elicitation, albeit with a narrower participation scope of 44 participants. Finally, Aakanksha et al. (2024) created a multi-lingual alignment dataset focused on local and global forms of harms by collecting 900 'harmful' prompts crafted by paid native-speakers in 8 languages.

Context Most of these studies are contextual in a way: they account for the impact of a labeller's context, such as language or religion, on their provided feedback. Aakanksha et al. (2024) further promotes contextualisation by instructing labellers to report and comment on local forms of harm. However, both Aakanksha et al. (2024) and Huang et al. (2024) offer limited background information about their participants beyond their native language or their location, respectively. Furthermore, only two of these studies delineate the contexts of specific applications: Rose, Kroner Dale, and Jusko (2025) delineate 10 use–cases, and Kirk et al. (2024b) employ topic clustering to identify the context of their recorded conversation. Conversely, Bergman et al. (2024), Aakanksha et al. (2024), and Huang et al. (2024) are eliciting guiding principles outside of specific contexts of use. Moreover, Huang et al. (2024) and Bergman et al. (2024) seek to create universal principles, yielding guidelines such as *"should not pretend to know the user personally"* (Bergman et al. 2024, p. 7), which may not suitably capture the nuances of varied use contexts.

Pluralism These works foster plurality by gathering input from participants with diverse demographics. Bergman et al. (2024) focuses on the importance of debate within communities by employing focus group methodologies in the phases of norms elicitation. However, most of these studies leave little room for the co-existence of conflicting perspectives and are instead consensus-driven. For instance, Bergman et al. (2024) and Huang et al. (2024) consolidate participant inputs via author–led rule harmonisation or voting to form a single set of general principles. While Aakanksha et al. (2024, p. 12031) acknowledges a "variation in the degree of harm" presented by annotators, they do not discuss the consequence of such plurality in their approach. The exception is Kirk et al. (2024b), which, while they highlight the potential of their dataset to identify consensus across opinion distributions, also highlight its applicability for personalised alignment, which could accommodate divergent viewpoints.

Participation All these studies are examples of participative approaches, as they each seek to include the voice of under-represented groups in alignment datasets by involving these groups in their creation. However, studies such as those by Aakanksha et al. (2024) and Huang et al. (2024) provide limited information regarding the working conditions of these annotators. In contrast, Bergman et al. (2024) details a participation format centred on small groups, emphasising a seemingly qualitative and well-remunerated experience for participants, who were given space for debate and self-expression. Notably, all these studies fall on the 'consultative' end of the participation framework (Delgado et al. 2023). Stakeholders provide input, but they are not empowered to influence the final configuration of language models. The limitation of this type of participation is illustrated by the outcome of Huang et al. (2024) effort: Anthropic adopted only one publicly voted principle in their subsequent LLMs despite acknowledging a low overlap between their own constitution and the publicly derived one (Anthropic 2024).

The Limits of Inclusion

The goal of these approaches is to build inclusive datasets that either refine rule sets or facilitate user personalisation. However, these efforts have been criticised for failing to genuinely empower marginalised groups. Young et al. (2024, p. 2) argue that public participation in consolidating rule sets ultimately bestows major industry actors with *"a degree of political legitimacy"* while enabling them to reduce safety

constraints to technical encodings within their *"centralised, powerful models"*. Similarly, Varshney (2024) contends that many commercial initiatives aimed at offering customisation are mostly superficial, being constrained within the moral frameworks of the companies: a form of 'pigeon-holing' that ensures these companies maintain ultimate control over the values propagated by their models. By retaining such control, these companies can impose their ways of knowing on their user base while suppressing other epistemologies, leading Varshney (2024) to characterise alignment, when mediated by metropole companies, as a reproduction of colonialist epistemic violence. Such violences are characterised by exclusionary practices – the systematic marginalisation and devaluation of alternative knowledge systems as a means of assimilation and the consolidation of dominance.

This is not to understate the value or relevance of these inclusion efforts. For instance, the PRISM datasets collected by Kirk et al. (2024b) provide crucial empirical insights into diverse and conflicting preferences, interaction types and priorities regarding LLMs. However, while inclusion is undoubtedly important, its ability to instigate meaningful change remains limited; inclusion mediated by those in power ultimately does not – and cannot – dislodge this power.

In response, Varshney (2024) proposes a shift toward concrete context-dependent and community-based normative alignment beyond inclusive datasets and universalism. He suggests employing Supervised Fine-Tuning (SFT) through Low-Rank Adaptation (LoRA) to create a repository of LoRA matrices – compressed representations of distinct community preferences – that can be applied to a base model at inference time (Hu et al. 2021).Varshney (2024)'s proposal offers a concrete strategy for transforming LLM alignment from a mechanism of institutional power consolidation into one of community empowerment, thus enabling pluralistic, participatory ownership alignment. However, his framework is presented as a one-size-fits-all solution, lacking specificity regarding the contexts in which it would be most effective, and omitting a discussion of circumstances where it may not be applicable. For example, Kirk et al. (2024a) point out the potential dangers of enabling such a personalisation of LLMs, such as polarisation arising from creating echo chambers or the essentialisation of communities through overly simplified and compressed representations. These risks would manifest at different levels and would necessitate different mitigation strategies depending on the application, as we will discuss in the next section. Therefore, a context-specific approach is essential both for identifying concrete methods to shift power back to impacted communities and for determining how this can be practically achieved.

Contextual, Pluralistic, and Participatory Alignment in Practice

Use Cases

To emphasise the importance of context in developing alignment strategies and to ground the discussion in concrete examples, two fictional use cases have been deliberately chosen for their stark contrast.

Voter Advice Applications (VAAs)

Background Voter Advice Applications (VAAs) are systems, such as chatbots, which help voters learn about upcoming elections or votations (Kamoen, McCartan, and Liebrecht 2022). Studies suggest that such systems can have positive effects on voter turnout, knowledge building and opinion formation, but may also influence final voting choices (Munzert and Ramirez-Ruiz 2021; Stadelmann-Steffen, Rajski, and Ruprecht 2023; Germann, Mendez, and Gemenis 2023).

Power Centralisation Recently, concerns have arisen regarding citizens' use of commercial LLM chatbots for voting information (Helming et al. 2023; Sharma, Liao, and Xiao 2024). Crucially, these systems are not specifically designed and tested to provide truthful and impartial voting recommendations. Furthermore, their control by non-democratic institutions shifts the power of informing and educating voters away from democratic institutions and political parties and toward private companies. This has led to increased anxiety about the impact of LLMs on the health of democratic processes (Coeckelbergh 2025). Indeed, this privatisation allows actors with vested interests to shape the global information landscape.

Alignment Mitigation To counter this power centralisation, democratic institutions or civil societies could develop tailored LLM systems for VAAs, offering a reliable alternative to commercial options. To guarantee the communication of up-to-date, verifiable information, these systems would likely be built using Retrieval-Augmented Generation (RAG), constraining answers to citations and summaries from defined, trusted sources. Alignment is crucial for ensuring that LLMs present retrieved information in a way that promotes dialogue and critical thinking. Sharma, Liao, and Xiao (2024) suggest that while LLM-powered conversational search can increase biased information querying and exacerbate existing biases compared to conventional web search, these issues are not insurmountable. They propose that the framing of the conversation can be designed to address these risks, for instance by highlighting the values of opposing arguments or identifying common ground. Therefore, in the context of VAAs, power redistribution could involve empowering local democratic institutions or trusted third parties to shape and oversee the information citizens receive, and the way that it is presented to them by LLM-based systems.

LLM-Powered Non-Playable Character (NPCs)

Background Non-playable characters (NPCs) are virtual characters within video games that players interact with but cannot control, ranging from allies and enemies to bystanders and pets. Traditionally, NPC interactions rely on scripted responses. However, the rise of large language models has led to the development of LLM-powered NPCs that offer free-form conversation capabilities, enhancing player immersion (Gallotta et al. 2024; Cox and Ooi 2024). While many small models could be deployed and aligned to the

conversational tone of different characters, this might not be a scalable solution as the number of side characters grows. Rather, a single model could be used to generate dialogue for all characters, meaning that the same model should be able to convincingly portray all characters simultaneously.

Power Centralisation While it's important for video games to feature a broad range of communities, respectful and accurate portrayals of marginalised groups often fall short because game studios lack diverse representation internally. This can lead to stereotypical depictions of communities, perpetuating colonial ideologies and erasing marginalised cultures, as seen in the misrepresentation of indigenous groups in games (Wallis 2023). LLMs also exhibit these harmful biases (Mitchell et al. 2025). Combining these two domains risks even more damaging portrayals. Therefore, to avoid harmful depictions, the power to shape representation needs to shift from private entities, game studios and LLM developers, directly to the communities being depicted.

Alignment Mitigation Rethinking alignment techniques is one way to make this power transfer happen. Diverse groups could be empowered to align models, creating more authentic representations that truly reflect their identities, values, and lived experiences. These community-trained models could then be licensed to game companies, allowing communities to 'rent out' their own representations and choose how they are depicted and in which games or contexts. This licensing model would also give them the power to rescind representation rights if they're uncomfortable with a storyline or character design, ensuring communities control not just *how* they are represented but also *which* games use their likenesses. However, such an approach necessitates alignment to be made accessible to a broad range of stakeholders.

Contextual Alignment

Contextual alignment means that every design decision in aligning language models must be sensitive to the specific context in which they operate. This sensitivity extends from the creation and curation of contextual alignment datasets to the strategic selection of appropriate alignment algorithms. In this section, we focus specifically on the latter, examining how algorithm choices should be driven by context. By demonstrating the contextual factors' impact on algorithm selection, we aim to illustrate that a one-size-fits-all approach cannot meet the decentralisation requirements of all use cases. Table 1 shows the contextual requirements for both use-cases, indicating the appropriate approaches needed in each case.

Consistent VAA There is evidence of a trade-off between the generalisation and diversity of alignment methods. Methods such as RLHF tend to experience greater 'mode collapse' than SFT. This means they will generate a smaller set of potential outputs to a given prompt, while simultaneously allowing for more robust maintenance of alignment guardrails across various use-cases (Kirk et al. 2024c; O'Mahony et al. 2024). For a VAA system, carefully tailored and nuanced formats of information presentation are crucial to avoid exacerbating bias, meaning that generalisation should be prioritised over creativity. In this context, RLHF's 'mode collapse' may be desirable as a means to enforce consistent responses across users. However, RLHF is associated with substantial implementation costs and complexities, making it potentially viable only for wealthier democratic societies. Moreover, iterative investigation is required to determine whether RLHF can provide sufficient guarantees of consistency and accuracy in such a sensitive use case.

Diverse NPCs In contrast, NPC use cases necessitate prioritising diversity over generalisation. Characters driven by the same LLM must generate varied dialogue reflective of their distinct personalities. While consistency in interactions and adherence to the storyline remains valuable, the priority is to create differing, engaging characters. Consequently, the reduction in diversity caused by RLHF makes it less suitable for this context, as it might cause very different characters to produce very similar dialogue, such as all of them repeating the same jokes (Jentzsch and Kersting 2023). Furthermore, the complexity of RLHF implementation would make it inaccessible to a broad range of communities. Instead, Parameter-Efficient Fine-Tuning (PEFT) methods would be preferable. These methods allow for fine-tuning a model by updating only a small subset of its parameters while achieving comparable performance to full SFT. Building on Varshney (2024)'s proposal, diverse communities can be empowered to create personalised representations of themselves, compressed into modular formats that can be applied to a common LLM. This concept can be realised through SFT combined with LoRA, a form of PEFT (Hu et al. 2021). LoRA matrices significantly reduce the computational resources needed for training and storage, as only the compact 'difference' (the LoRA matrix) between the base model and the community's preferences is stored and applied during inference. In this framework, a specific representation tailored to a character's attributes can be developed, enabling communities to train and control how they are portrayed. Realising this would likely necessitate communities either directly labelling datasets to encode their specific preferences or inferring these preferences from existing unstructured textual data they have produced (Padhi et al. 2024).

Rigorous testing is essential to determine whether diverse communities can satisfactorily use LoRA training to generate representations that they find acceptable and whether it compensates for biases present in the pre-trained model. While identities are inherently complex, fluid, and dynamic, making their complete capture by a machine learning system unfeasible, the goal of this use case is not to replicate individuals perfectly. Instead, it aims to develop a system that captures their likeness, ensuring that their representation within the game's universe is both respectful and appreciated by the community.

Pluralistic Alignment

Sorensen et al. (2024) propose several approaches to mitigate the struggle of current alignment approaches with pluralism. Two particularly relevant approaches here are Over-

	Voter Advice Applications (VAAs)	Non-Playable Characters (NPCs)
Stakes	Integrity of public debate and democratic processes	Fair community representation, preventing harmful stereotypes and digital erasure
Power Imbalance	Between nations, or between private entities and democratic states	Between game development companies and depicted communities
Contextual Alignment	*RLHF* for consistency and reliability	*LoRA* for accessibility and diversity
Pluralism Approach	*Overton* to represent a spectrum of political perspectives and mitigate voter polarisation.	*Steerable* for dialogue consistency with character narratives and community values
Participation Model	*Centralised*, state-level participation, consensus seeking among political parties and citizen representatives	*Decentralized*, community-level participation, self-organized, no inter-community consensus requirement

Table 1: Comparative alignment characteristics of Voter Advice Applications and Non-Playable Characters, highlighting the need for use-case grounded approaches to contextual, pluralistic, and participative alignment.

ton pluralism and steerable pluralism. Overton pluralistic systems answer a question with a spectrum of reasonable responses. They are crucial when addressing topics with multiple, valid perspectives. A steerable pluralistic system, on the other hand, is one which *"faithfully steers its response to a given attribute or perspective"* (Sorensen et al. 2024, p. 3). These steering attributes could be features such as 'political leaning', 'religion', 'age', or 'nationality'. In this approach, only the point of view that best suits the steering attributes is presented. Other approaches, such as distributional pluralism, suggest that the likelihood of a viewpoint being presented should match its prevalence in human populations. While this may work for political simulations (Yang et al. 2025), it is unsuitable for VAAs, which should always encompass a range of possible views, and NPCs, which must represent only the intended perspective.

Overton Pluralistic VAA Overton pluralism is likely the most suited form of pluralism for VAAs. As previously discussed, models should be aligned to present retrieved information in a carefully designed format, such as highlighting the values of opposing arguments or emphasising common ground. 'Reasonable' answers could be extracted from a predefined source of truth, curated by democratic institutions, trusted third parties, or a consortium of political parties. These sources could also be accompanied by a polling of consensus across political parties, allowing models to incorporate such information into the weighting of outputs, similar to how current VAAs operate (e.g., Politools 2019). Capturing the agreement, divergence and uncertainty of different political parties could also be achieved by adapting the methodologies proposed by Kuo et al. (2024). Their approaches would allow political annotators to continuously curate the VAA's database, highlighting ambiguities and differences through their discussion. These discussions can, in turn, be fed into the model to guide the presentation of information.

Another way to implement pluralism in VAA would be to personalise advice and voting recommendations based on the political leaning of a user through steerable pluralism. However, this could be a misguided approach as it could exacerbate the concerning trend of voter polarisation by creating echo chambers similar to those produced by pro-attitudinal media (Kirk et al. 2024a).

Steerable Pluralistic NPCs In contrast, an NPCs dialogue should be tailored to the character's personality, instead of being representative of all characters' diversity at once. Therefore, NPCs should be steerable pluralistic, where steering attributes, such as identity or character traits, are used to index a corresponding LoRA matrix. This selected matrix is then applied to the base model to produce the desired representation. However, such a process risks relying on rigid categorisations that would clash with the fluidity of the categories they aim to reflect (Suresh et al. 2022). The challenge lies in defining the attribute set without oversimplification or harmful stereotypes. Maintaining nuance and complexity in the representation of complex identities calls for an intersectional approach (Sorensen et al. 2024). The intersections of each social identity would likely generate new, unique attributes (Crenshaw 1998). For example, the intersection of being elderly and from a low socioeconomic background might create a new 'wary of authority' attribute. Designing nuanced and intersectional attributes requires careful work and the involvement of the groups they represent. In theory, it might be possible to generate intersectional representations by combining LoRA matrices trained on individual attributes, such as 'elderly' and 'working class', with the hope of synthesising an attribute like 'wary of authority' (Zhang et al. 2023). However, this method would not necessarily yield meaningful and respectful intersectional representations. Rather, such automation risks marginalising and excluding the very communities it aims to represent. For instance, the presumption that intersectional attributes can be synthesised from individual components might justify the exclusion of elderly working-class groups from the process entirely. Instead, intersectional communities must be actively involved in the creation of their own representations, granted that they wish to be portrayed at all. Therefore, a more meaningful approach to steerable alignment using LoRA could involve creating accessible infrastructure for communities to provide feedback on various LLM outputs, such as the LLM Arena (Chiang et al. 2024). This infrastructure could include tools that utilise this feedback to generate a representative matrix. Such tools could be developed by universities

or funded by public resources.

Participatory Alignment

Participation can range from centralised and consultative forms to decentralised formats, which facilitate full ownership. None of these approaches is intrinsically better; rather, each is better suited to specific use cases, depending on the underlying conceptualisation of power redistribution.

Centrally Designed VAAs Decentralising power, in the context of VAA, entails resisting the power shift to inform citizens toward private, often US or China-based, companies, to local democratic institutions. However, LLMs systems used to disseminate political information might still need some centralised oversight at the level of local states or through third-party institutions. Indeed, decentralisation at a political party or citizen level might prove unhealthy. For example, allowing each political party to deploy its own model for informing citizens may risk creating dangerous echo chambers and citizen polarisation. Nonetheless, participation is essential to prevent any single institution from dominating the political information landscape. Political parties, citizens with varying degrees of political literacy, and independent monitoring committees should all be involved in the creation and alignment of a political LLM. While ensuring meaningful agency for all participants in the design process is crucial, achieving a workable consensus remains necessary for determining the effective targets of alignment. In cases where fundamental disagreements or resource constraints make full consensus infeasible, a centralised authority may be required to establish a practical resolution (Pols and Spahn 2015).

Therefore, participation can help prevent power from becoming concentrated in the hands of foreign companies and incorporate the perspectives of local political parties and citizens. However, full transfer of ownership to the participants, be it political parties or citizens, would not be optimal. Some level of centralised oversight and authority might be required to deploy such a system in practice. Such participation and alignment efforts are likely to be expensive. Still, they could potentially be financed through the same processes that finance current VAA system development, such as federal agencies, media companies, or non-profit organisations (Stockinger et al. 2024).

Community-led NPCs Conversely, decisions regarding the appropriate alignment of NPCs with specific communities do not need global consensus. In fact, each community can independently determine its own alignment, without requiring consultation or agreement with other groups. The granularity of these communities can be flexibly defined, allowing subgroups to diverge. For instance, if a faction of bankers disagrees with the alignment representation established by the broader banking community, they are free to form a distinct subgroup, such as 'bankers for Bitcoin', and develop their own tailored representations. A centralised alternative, where game studios organise participatory processes and invite community involvement to ensure satisfactory representation, risks creating forms of 'participation-washing' (Sloane et al. 2022). In such scenarios, communities might be exploited to generate inexpensive data under the guise of inclusion, ultimately increasing game studio profit margins without granting communities concrete power, such as the ability to retract their participation. Genuinely transferring power to shape representation back to communities might be better achieved by community-led and -owned processes. In this model, communities align their own models, either on pre-existing data or on labels they generate themselves, and share the resulting licensed LoRA matrix instead of their raw data. Birhane et al. (2022) illustrate a similar process as a powerful mechanism for reciprocity and refusal through the example of the Māori community's development and subsequent licensing of self-representative data, which empowered them to determine who may utilise their data and under what parameters. Consequently, a fully decentralised form of participation, granting communities complete oversight and ownership of their aligned models, offers a compelling approach to re-establish community control over self-representation. While game companies would retain control over storylines and overarching dialogue, the specific format and style of character portrayal would be dictated by the community's alignment. Moreover, if a community deems a storyline insensitive or disrespectful, it would have the power to revoke licensing rights, ensuring accountability.

However, this approach is not without its limitations. Not all communities possess the necessary resources, time, or technical expertise to train a LoRA matrix. Significant investment in infrastructure is essential to enable meaningful participation in large-scale machine learning systems (Young et al. 2024). Moreover, this approach risks a 'naive' approach to cultural relativism (Gabriel 2020), where all moral beliefs are framed as equally valid within their respective contexts, which has the potential to be weaponised to justify oppression in the name of cultural values. For example, communities advocating violence against others should arguably not be provided a platform to disseminate such ideas. Game studios may need to develop content moderation infrastructures to manage and oversee the outputs of their deployed models.

Therefore, the centrally organised participation appropriate for aligning a VAA system would likely not be suitable for NPCs alignment. Redistributing the power to shape narratives and representations of communities would be most effectively achieved through decentralised and community-owned forms of participation.

Conclusion

Alignment can serve as a meaningful site of resistance against the current centralisation of power in knowledge replication and distribution enabled by Large Language Models. By emphasising a contextual, pluralistic, and participatory alignment process, we outline not a rigid formula but a set of guiding principles that must be tailored to the specific systems they intend to enhance. These conditions serve as a roadmap for reimagining alignment systems in a way that redistributes epistemic power and fosters greater democratic participation in knowledge creation. Moving forward,

it is essential to implement and empirically test these alignment strategies, carefully measuring how effectively communities can assert control over the knowledge generated by LLMs, and identifying the extent to which this power remains consolidated among the original developers. Furthermore, questions around how these efforts are organised in practice and how they are financed remain and are likely to be limiting factors in the decentralisation of LLM alignment.

While broader societal transformations are undoubtedly necessary for achieving genuine epistemic justice, rethinking LLM alignment approaches can address immediate risks posed by contemporary LLM systems, such as the erosion of democratic processes.

Acknowledgments

The authors would like to thank Dr. Dorian Peters and Professor Elena Simperl for their feedback. This work was supported by the Engineering and Physical Sciences Research Council, grant number EP/Y009800/1, Responsible AI UK.

References

Aakanksha; Ahmadian, A.; Ermis, B.; Goldfarb-Tarrant, S.; Kreutzer, J.; Fadaee, M.; and Hooker, S. 2024. The Multilingual Alignment Prism: Aligning Global and Local Preferences to Reduce Harm. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 12027–12049. Miami, Florida, USA: Association for Computational Linguistics.

Anthropic. 2023. Collective Constitutional AI: Aligning a Language Model with Public Input.

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. Technical report, Anthropic.

Atari, M.; Xue, M. J.; Park, P. S.; Blasi, D.; and Henrich, J. 2023. Which Humans?

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; Chen, C.; Olsson, C.; Olah, C.; Hernandez, D.; Drain, D.; Ganguli, D.; Li, D.; Tran-Johnson, E.; Perez, E.; Kerr, J.; Mueller, J.; Ladish, J.; Landau, J.; Ndousse, K.; Lukosuite, K.; Lovitt, L.; Sellitto, M.; Elhage, N.; Schiefer, N.; Mercado, N.; DasSarma, N.; Lasenby, R.; Larson, R.; Ringer, S.; Johnston, S.; Kravec, S.; Showk, S. E.; Fort, S.; Lanham, T.; Telleen-Lawton, T.; Conerly, T.; Henighan, T.; Hume, T.; Bowman, S. R.; Hatfield-Dodds, Z.; Mann, B.; Amodei, D.; Joseph, N.; McCandlish, S.; Brown, T.; and Kaplan, J. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.

Bergman, S.; Marchal, N.; Mellor, J.; Mohamed, S.; Gabriel, I.; and Isaac, W. 2024. STELA: A Community-Centred Approach to Norm Elicitation for AI Alignment. *Scientific Reports*, 14(1): 6616.

Birhane, A.; Isaac, W.; Prabhakaran, V.; Diaz, M.; Elish, M. C.; Gabriel, I.; and Mohamed, S. 2022. Power to the People? Opportunities and Challenges for Participatory AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–8. Arlington VA USA: ACM. ISBN 978-1-4503-9477-2.

Blodgett, S. L.; Barocas, S.; Daumé III, H.; and Wallach, H. 2020. Language (Technology) Is Power: A Critical Survey of "Bias" in NLP. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476. Online: Association for Computational Linguistics.

Bommasani, R.; Creel, K. A.; Kumar, A.; Jurafsky, D.; and Liang, P. 2022. Picking on the Same Person: Does Algorithmic Monoculture Lead to Outcome Homogenization? In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, 3663–3678. Red Hook, NY, USA: Curran Associates Inc. ISBN 978-1-7138-7108-8.

Casper, S.; Davies, X.; Shi, C.; Gilbert, T. K.; Scheurer, J.; Rando, J.; Freedman, R.; Korbak, T.; Lindner, D.; Freire, P.; Wang, T. T.; Marks, S.; Segerie, C.-R.; Carroll, M.; Peng, A.; Christoffersen, P.; Damani, M.; Slocum, S.; Anwar, U.; Siththaranjan, A.; Nadeau, M.; Michaud, E. J.; Pfau, J.; Krasheninnikov, D.; Chen, X.; Langosco, L.; Hase, P.; Biyik, E.; Dragan, A.; Krueger, D.; Sadigh, D.; and Hadfield-Menell, D. 2023. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research*.

Cave, S.; and Dihal, K. 2020. The Whiteness of AI. *Philosophy & Technology*, 33(4): 685–703.

Chen, Z.; Cano, A. H.; Romanou, A.; Bonnet, A.; Matoba, K.; Salvi, F.; Pagliardini, M.; Fan, S.; Köpf, A.; Mohtashami, A.; Sallinen, A.; Sakhaeirad, A.; Swamy, V.; Krawczuk, I.; Bayazit, D.; Marmet, A.; Montariol, S.; Hartley, M.-A.; Jaggi, M.; and Bosselut, A. 2023. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. arXiv:2311.16079.

Chiang, W.-L.; Zheng, L.; Sheng, Y.; Angelopoulos, A. N.; Li, T.; Li, D.; Zhu, B.; Zhang, H.; Jordan, M. I.; Gonzalez, J. E.; and Stoica, I. 2024. Chatbot arena: an open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, 8359–8388. Vienna, Austria: JMLR.org.

Coeckelbergh, M. 2025. LLMs, Truth, and Democracy: An Overview of Risks. *Science and Engineering Ethics*, 31(1): 4.

Cox, S. R.; and Ooi, W. T. 2024. Conversational Interactions with NPCs in LLM-Driven Gaming: Guidelines from a Content Analysis of Player Feedback. In Følstad, A.; Araujo, T.; Papadopoulos, S.; Law, E. L.-C.; Luger, E.; Goodwin, M.; Hobert, S.; and Brandtzaeg, P. B., eds., *Chatbot Research and Design*, 167–184. Cham: Springer Nature Switzerland. ISBN 978-3-031-54975-5.

Crenshaw, K. 1998. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory, and Antiracist Politics. In Phillips, A., ed., *Feminism And Politics: Oxford Readings In Feminism*. Oxford University Press. ISBN 978-0-19-878206-3.

Delgado, F.; Yang, S.; Madaio, M.; and Yang, Q. 2023. The Participatory Turn in AI Design: Theoretical Founda-

tions and the Current State of Practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, 1–23. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0381-2.

Diaz, M.; Kivlichan, I. D.; Rosen, R.; Baker, D. K.; Amironesei, R.; Prabhakaran, V.; and Denton, E. 2022. CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation. In *2022 ACM Conference on Fairness Accountability and Transparency*, 2342–2351.

Durmus, E.; Nyugen, K.; Liao, T.; Schiefer, N.; Askell, A.; Bakhtin, A.; Chen, C.; Hatfield-Dodds, Z.; Hernandez, D.; Joseph, N.; Lovitt, L.; McCandlish, S.; Sikder, O.; Tamkin, A.; Thamkul, J.; Kaplan, J.; Clark, J.; and Ganguli, D. 2023. Towards Measuring the Representation of Subjective Global Opinions in Language Models. *ArXiv*.

Foucault, M. 1975. *Surveiller et Punir: Naissance de Ia Prison*. Editions Gallimard,Paris.

Gabriel, I. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3): 411–437.

Gallotta, R.; Todd, G.; Zammit, M.; Earle, S.; Liapis, A.; Togelius, J.; and Yannakakis, G. N. 2024. Large Language Models and Games: A Survey and Roadmap.

Ganesh, M. I.; and Moss, E. 2022. Resistance and Refusal to Algorithmic Harms: Varieties of 'Knowledge Projects'. *Media International Australia*, 183(1): 90–106.

Gebru, T.; and Torres, É. P. 2024. The TESCREAL Bundle: Eugenics and the Promise of Utopia through Artificial General Intelligence. *First Monday*.

Germann, M.; Mendez, F.; and Gemenis, K. 2023. Do Voting Advice Applications Affect Party Preferences? Evidence from Field Experiments in Five European Countries. *Political Communication*, 40(5): 596–614.

Gillespie, T. 2024. Generative AI and the Politics of Visibility. *Big Data & Society*, 11(2): 20539517241252131.

Gray, M. L.; and Suri, S. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston New York NY: Harper Business, illustrated edition edition. ISBN 978-1-328-56624-9.

Heaven, W. D. 2023. The inside Story of How ChatGPT Was Built from the People Who Made It. *MIT Technology Review*.

Helming, C.; Müller, A.; Spielkamp, M.; Schiller, A. L.; Kesler, W.; Omalar, M.; Thümmler, M.; Zimmermann, M.; Sanchez, I.; Kimel, A.; Pannatier, E.; Urech, T.; Sorie, D.; Loi, M.; Felder, A.; Romano, S.; Kerby, N.; Angius, R.; Robutti, S.; Schueler, M.; Faddoul, M.; and Çetin, R. B. 2023. Generative AI and Elections: Are Chatbots a Reliable Source of Information for Voters? Technical report, AI Forensic.

Henrich, J.; Heine, S. J.; and Norenzayan, A. 2010. The Weirdest People in the World? *Behavioral and Brain Sciences*, 33(2-3): 61–83.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Huang, S.; Siddarth, D.; Lovitt, L.; Liao, T. I.; Durmus, E.; Tamkin, A.; and Ganguli, D. 2024. Collective Constitutional AI: Aligning a Language Model with Public Input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1395–1417. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0450-5.

Jakesch, M.; Bhat, A.; Buschek, D.; Zalmanson, L.; and Naaman, M. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, 1–15. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9421-5.

Jasanoff, S. 2006. *States of Knowledge: The Co-Production of Science and the Social Order*. London: Routledge, 1st edition edition. ISBN 978-0-415-40329-0.

Jentzsch, S. F.; and Kersting, K. 2023. ChatGPT Is Fun, but It Is Not Funny! Humor Is Still Challenging Large Language Models. In *61st Annual Meeting of the Association for Computational Linguistics, ACL 2023*. Toronto, Canada. ISBN 978-1-959429-87-6.

Jha, P.; Jain, R.; Mandal, K.; Chadha, A.; Saha, S.; and Bhattacharyya, P. 2024. MemeGuard: An LLM and VLM-based Framework for Advancing Content Moderation via Meme Intervention. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8084–8104. Bangkok, Thailand: Association for Computational Linguistics.

Kamoen, N.; McCartan, T.; and Liebrecht, C. 2022. Conversational Agent Voting Advice Applications: A Comparison Between a Structured, Semi-structured, and Non-structured Chatbot Design for Communicating with Voters About Political Issues. In Følstad, A.; Araujo, T.; Papadopoulos, S.; Law, E. L.-C.; Luger, E.; Goodwin, M.; and Brandtzaeg, P. B., eds., *Chatbot Research and Design*, 160–175. Cham: Springer International Publishing. ISBN 978-3-030-94890-0.

Katzman, J.; Wang, A.; Scheuerman, M.; Blodgett, S. L.; Laird, K.; Wallach, H.; and Barocas, S. 2023. Taxonomizing and Measuring Representational Harms: A Look at Image Tagging. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, volume 37 of *AAAI'23/IAAI'23/EAAI'23*, 14277–14285. AAAI Press. ISBN 978-1-57735-880-0.

Kirk, H.; Vidgen, B.; Rottger, P.; and Hale, S. 2023. The Empty Signifier Problem: Towards Clearer Paradigms for Operationalising "Alignment" in Large Language Models. In *Socially Responsible Language Modelling Research*.

Kirk, H. R.; Vidgen, B.; Röttger, P.; and Hale, S. A. 2024a. The Benefits, Risks and Bounds of Personalizing the Align-

ment of Large Language Models to Individuals. *Nature Machine Intelligence*, 6(4): 383–392.

Kirk, H. R.; Whitefield, A.; Röttger, P.; Bean, A. M.; Margatina, K.; Mosquera, R.; Ciro, J. M.; Bartolo, M.; Williams, A.; He, H.; Vidgen, B.; and Hale, S. A. 2024b. The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Kirk, R.; Mediratta, I.; Nalmpantis, C.; Luketina, J.; Hambro, E.; Grefenstette, E.; and Raileanu, R. 2024c. Understanding the Effects of RLHF on LLM Generalisation and Diversity. *The Twelfth International Conference on Learning Representations*.

Kniaz, R. 2023. The Incoming Tidal Wave Of Data Pollution In AI. *Forbes*.

Kuo, T.-S.; Halfaker, A.; Cheng, Z.; Kim, J.; Wu, M.-H.; Wu, T.; Holstein, K.; and Zhu, H. 2024. Wikibench: Community-Driven Data Curation for AI Evaluation on Wikipedia. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–24.

Latour, B. 1988. *Science in Action: How to Follow Scientists and Engineers Through Society*. Cambridge (Mass.): Harvard University Press, revised ed. edition edition. ISBN 978-0-674-79291-3.

Le Ludec, C.; Cornet, M.; and Casilli, A. A. 2023. The Problem with Annotation. Human Labour and Outsourcing between France and Madagascar. *Big Data & Society*, 10(2): 20539517231188723.

Lorde, A. 2007. *Sister Outsider: Essays and Speeches*. [The Crossing Press Feminist Series]. Berkeley [California: Crossing Press, revised edition. edition. ISBN 978-1-58091-186-3.

Mamykina, L.; Manoim, B.; Mittal, M.; Hripcsak, G.; and Hartmann, B. 2011. Design Lessons from the Fastest Q&a Site in the West. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, 2857–2866. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-0228-9.

Miceli, M.; and Posada, J. 2022. The Data-Production Dispositif. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 460:1–460:37.

Miceli, M.; Schuessler, M.; and Yang, T. 2020. Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2): 115:1–115:25.

Mitchell, M.; Attanasio, G.; Baldini, I.; Clinciu, M.; Clive, J.; Delobelle, P.; Dey, M.; Hamilton, S.; Dill, T.; Doughman, J.; Dutt, R.; Ghosh, A.; Forde, J. Z.; Holtermann, C.; Kaffee, L.-A.; Laud, T.; Lauscher, A.; Lopez-Davila, R. L.; Masoud, M.; Nangia, N.; Ovalle, A.; Pistilli, G.; Radev, D.; Savoldi, B.; Raheja, V.; Qin, J.; Ploeger, E.; Subramonian, A.; Dhole, K.; Sun, K.; Djanibekov, A.; Mansurov, J.; Yin, K.; Cueva, E. V.; Mukherjee, S.; Huang, J.; Shen, X.; Gala, J.; Al-Ali, H.; Djanibekov, T.; Mukhituly, N.; Nie, S.; Sharma, S.; Stanczak, K.; Szczechla, E.; Timponi Torrent, T.; Tunuguntla, D.; Viridiano, M.; Van Der Wal, O.; Yakefu, A.; Névéol, A.; Zhang, M.; Zink, S.; and Talat, Z. 2025. SHADES: Towards a Multilingual Assessment of Stereotypes in Large Language Models. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 11995–12041. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-189-6.

Mittelstadt, B. 2019. Principles Alone Cannot Guarantee Ethical AI. *Nature Machine Intelligence*, 1(11): 501–507.

Mohamed, S.; Png, M.-T.; and Isaac, W. 2020. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology*, 33(4): 659–684.

Muldoon, J.; Cant, C.; Graham, M.; and Ustek Spilda, F. 2023. The Poverty of Ethical AI: Impact Sourcing and AI Supply Chains. *AI & SOCIETY*.

Muldoon, J.; and Wu, B. A. 2023. Artificial Intelligence in the Colonial Matrix of Power. *Philosophy & Technology*, 36(4): 80.

Munzert, S.; and Ramirez-Ruiz, S. 2021. Meta-Analysis of the Effects of Voting Advice Applications. *Political Communication*.

O'Mahony, L.; Grinsztajn, L.; Schoelkopf, H.; and Biderman, S. 2024. Attributing Mode Collapse in the Fine-Tuning of Large Language Models. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training Language Models to Follow Instructions with Human Feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, 27730–27744. Red Hook, NY, USA: Curran Associates Inc. ISBN 978-1-7138-7108-8.

Padhi, I.; Natesan Ramamurthy, K.; Sattigeri, P.; Nagireddy, M.; Dognin, P.; and Varshney, K. R. 2024. Value Alignment from Unstructured Text. In Dernoncourt, F.; Preoţiuc-Pietro, D.; and Shimorina, A., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 1083–1095. Miami, Florida, US: Association for Computational Linguistics.

Politools. 2019. Methodenbeschreibung – Smartvote Wahlempfehlung. Technical report, Politools.

Pols, A.; and Spahn, A. 2015. Design for the Values of DemocracyDemocracyand JusticeJustice. In van den Hoven, J.; Vermaas, P. E.; and van de Poel, I., eds., *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*, 335–363. Dordrecht: Springer Netherlands. ISBN 978-94-007-6970-0.

Qadri, R.; Diaz, M.; Wang, D.; and Madaio, M. 2025. The Case for "Thick Evaluations" of Cultural Representation in AI. arXiv:2503.19075.

Raji, D.; Denton, E.; Bender, E. M.; Hanna, A.; and Paullada, A. 2021. AI and the Everything in the Whole Wide World Benchmark. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1.

Rose, K.; Kroner Dale, M.; and Jusko, K. 2025. From Community Input to AI Alignment: Incorporating Global Perspectives in AI Development. In *Participatory AI Research & Practice Symposium*. Paris.

Sharma, N.; Liao, Q. V.; and Xiao, Z. 2024. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, 1–17. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0330-0.

Sloane, M.; Moss, E.; Awomolo, O.; and Forlano, L. 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*, EAAMO '22, 1–6. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9477-2.

Sorensen, T.; Moore, J.; Fisher, J.; Gordon, M. L.; Mireshghallah, N.; Rytting, C. M.; Ye, A.; Jiang, L.; Lu, X.; Dziri, N.; Althoff, T.; and Choi, Y. 2024. Position: A Roadmap to Pluralistic Alignment. In *Proceedings of the 41st International Conference on Machine Learning*, 46280–46302. PMLR.

Stadelmann-Steffen, I.; Rajski, H.; and Ruprecht, S. 2023. The Role of Vote Advice Application in Direct-Democratic Opinion Formation: An Experiment from Switzerland. *Acta Politica*, 58(4): 792–818.

Stockinger, E.; Maas, J.; Talvitie, C.; and Dignum, V. 2024. Trustworthiness of voting advice applications in Europe. *Ethics and Information Technology*, 26(3): 55.

Suchman, L.; and Suchman, L. A. 2007. *Human-Machine Reconfigurations: Plans and Situated Actions*. Cambridge University Press. ISBN 978-0-521-67588-8.

Suresh, H.; Movva, R.; Dogan, A. L.; Bhargava, R.; Cruxen, I.; Cuba, A. M.; Taurino, G.; So, W.; and D'Ignazio, C. 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*, FAccT '22, 667–678. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9352-2.

Suresh, H.; Tseng, E.; Young, M.; Gray, M.; Pierson, E.; and Levy, K. 2024. Participation in the age of foundation models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1609–1621. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0450-5.

Varshney, K. R. 2024. Decolonial AI Alignment: Openness, Visesa-Dharma, and Including Excluded Knowledges. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 1467–1481.

Victor, D. 2016. Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk. *The New York Times*.

Wallis, D. 2023. Appropriation or Erasure? Imagining Indigenous Futures in Games. *Journal of Games Criticism*, Volume 5(Bonus Issue A).

Wang, D.; Prabhat, S.; and Sambasivan, N. 2022. Whose AI Dream? In Search of the Aspiration in Data Annotation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, 1–16. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9157-3.

Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.-S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; Kenton, Z.; Brown, S.; Hawkins, W.; Stepleton, T.; Biles, C.; Birhane, A.; Haas, J.; Rimell, L.; Hendricks, L. A.; Isaac, W.; Legassick, S.; Irving, G.; and Gabriel, I. 2021. Ethical and Social Risks of Harm from Language Models.

Weidinger, L.; Uesato, J.; Rauh, M.; Griffin, C.; Huang, P.-S.; Mellor, J.; Glaese, A.; Cheng, M.; Balle, B.; Kasirzadeh, A.; Biles, C.; Brown, S.; Kenton, Z.; Hawkins, W.; Stepleton, T.; Birhane, A.; Hendricks, L. A.; Rimell, L.; Isaac, W.; Haas, J.; Legassick, S.; Irving, G.; and Gabriel, I. 2022. Taxonomy of Risks Posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 214–229. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9352-2.

Yang, J. C.; Dailisan, D.; Korecki, M.; Hausladen, C. I.; and Helbing, D. 2025. LLM Voting: Human Choices and AI Collective Decision-Making. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '24, 1696–1708. San Jose, California, USA: AAAI Press.

Young, M.; Ehsan, U.; Singh, R.; Tafesse, E.; Gilman, M.; Harrington, C.; and Metcalf, J. 2024. Participation versus Scale: Tensions in the Practical Demands on Participatory AI. *First Monday*.

Zhang, J.; Chen, S.; Liu, J.; and He, J. 2023. Composing Parameter-Efficient Modules with Arithmetic Operation. *Advances in Neural Information Processing Systems*, 36: 12589–12610.