Aligning AI with Public Values: Deliberation and Decision-Making for Governing Multimodal LLMs in Political Video Analysis

Tanusree Sharma[1], **Yujin Potter**[3], **Zachary Kilhoffer**[2], **Yun Huang**[2], **Dawn Song**[3], **Yang Wang** [2]
[1] Pennsylvania State University
[2] University of Illinois at Urbana Champaign
[3] University of California, Berkeley
[1] tanusree.sharma@psu.edu

Abstract

How AI models should deal with political topics has been discussed, but it remains challenging and requires better governance. This paper examines the governance of large language models through individual and collective deliberation, focusing on politically sensitive videos. We conducted a two-step study: interviews with 10 journalists established a baseline understanding of expert video interpretation; 114 individuals through deliberation using *Inclusive.AI*, a platform that facilitates democratic decision-making through decentralized autonomous organization (DAO) mechanisms. Our findings reveal distinct differences in interpretative priorities: while experts emphasized emotion and narrative, general public prioritized factual clarity, objectivity, and emotional neutrality. Furthermore, we examined how different governance mechanisms - quadratic vs. weighted voting and equal vs. 20/80 voting power - shape users' decision-making regarding AI behavior. Results indicate that voting methods significantly influence outcomes, with quadratic voting reinforcing perceptions of liberal democracy and political equality. Our study underscores the necessity of selecting appropriate governance mechanisms to better capture user perspectives and suggests decentralized AI governance as a potential way to facilitate broader public engagement in AI development, ensuring that varied perspectives meaningfully inform design decisions.

Introduction

A major criticism of AI development is the lack of transparency, particularly the insufficient documentation, and traceability in model design, specification, and deployment (Brundage et al. 2020), leading to adverse outcomes including discrimination, lack of representation, and breaches of legal regulations. Traditional social science approaches, such as interviews and surveys, often fall short in capturing user expectations due to their limitations in facilitating ongoing deliberation. Governance, in contrast, is an interdisciplinary research area that involves stakeholders, (Shneiderman 2020a; Bu et al. 2020; Rubinstein and Good 2013; Wang, Hayes, and Bashir 2022) for structural changes, such as defining bias criteria, determining rules for dataset diversity, etc. This involves principles such as normative positions, concrete actions, and engineering practices.

AI governance literature often clusters into key themes, many borrowed from data protection and privacy fields- (1) FACT - fairness, accuracy, confidentiality, and transparency (Kemper and Kolkman 2019; Kaminski and Malgieri 2020; Selbst 2021); (2) FATE - fairness, accountability, transparency, and ethics (Barocas, Hood, and Ziewitz 2013); (3) privacy preservation; (4) governance, compliance, and risk (Calo 2017; Gasser and Almeida 2017; Scherer 2015; Butcher and Beridze 2019); (5) trust and safety (Biden 2023; Shneiderman 2020b; Wang, Hayes, and Bashir 2022; Saravanakumar and Arun 2014; Biden 2023); and (6) alignment with human values (Ji et al. 2024; Norhashim and Hahn 2024). Additionally, there is a growing focus on participatory AI (Young et al. 2024) leveraging existing international legal frameworks (Cihon 2019; Maas 2021; Wallach and Marchant 2018; Erdélyi and Goldsmith 2018). The AI Executive Order further highlights the need for a coordinated approach, emphasizing community engagement (Biden 2023).

Emerging models such as Decentralized Autonomous Organizations (DAOs) (Sharma et al. 2023a) also provide innovative directions for technical elements that support varied structural concepts from management science and community coordination. DAOs are blockchain-based organizations governed by smart contracts and decentralized decision-making, enabling collective governance without centralized control (Sharma et al. 2023a). By leveraging transparent, automated processes with smart contract governance, DAO provides a potential empirical testbed for exploring social choice experiments in potentially improving the current AI governance structure through a computational lens (Benkler, Shaw, and Hill 2015; Lalley and Weyl 2018; Weyl, Ohlhaver, and Buterin 2022; Zhang and Zhou 2017; Weber 2015). However, a fundamental tension exists between participatory decision-making in AI and its global, distributed nature (Young et al. 2024). DAOs present unique opportunities to address this challenge by implementing mechanisms such as social choice designs, quadratic voting, and liquid democracy (Lalley and Weyl 2018; Weyl, Ohlhaver, and Buterin 2022; Zhang and Zhou 2017), while also enabling anonymous participation for diverse voices.

To examine the benefits of decentralized governance in AI development, we conduct a case study focusing on how AI systems should address politically sensitive topics. The use of LLMs in political domains has been widely debated, in-

cluding their political biases (Potter et al. 2024a,b; Rozado 2024; Feng et al. 2023; Santurkar et al. 2023). Recent studies have revealed that LLMs can influence users' political views through their interactions (Potter et al. 2024b; Fisher et al. 2024; Costello, Pennycook, and Rand 2024). While several approaches have been proposed to pursue the political neutrality of LLMs, no clear consensus has emerged (Potter et al. 2024a); for example, many users expressed enjoyment when they are engaged in the interaction with politically leaned LLMs (Potter et al. 2024b). The conflicting views on these issues highlight the need for a deliberative process to incorporate diverse user perspectives. This motivates our research questions:

RQ1: How does the general public perceive the use of LLM in political content interpretation?

RQ2: How do DAO governance mechanisms influence public opinions about improving LLM design?

We propose **Inclusive.AI**, a DAO-enabled governance, emphasizing inclusivity and human oversight in LLM design oversight. As illustrated in Figure 1, to explicitly understand users' specific expectations, the governance model allows users to deliberate on sensitive topics where LLM output can be controversial and contentious. For our experiment, we used a video from the 2020 US presidential debate as a case study to explore public preferences in governing LLM behavior (Linegar, Kocielnik, and Alvarez 2023). To ensure secure and equitable participation, we implemented DAO infrastructure to enhance trust in the governance process. With Inclusive.AI, users first deliberate on LLM outputs, express their preferences and then participate in governance voting to guide future LLM design for political video interpretations.

Findings. Through an online experiment of 114 US internet users, our findings highlighted overlapping values between individual and collective deliberation for improving LLM output for political video content. Some factors are considered important, including, the emotions of the speaker, subjective content (e.g., who supports or opposes, composure, professionalism), and the speaker's positionality. There are some distinct differences in interpretative priorities: while experts emphasized emotion and narrative, general public prioritized factual clarity, objectivity, and emotional neutrality. Our findings also highlighted participants' perceived quality of the governance of the Inclusive.AI tool whereas voting methods significantly influence outcomes, with quadratic voting reinforcing perceptions of liberal democracy and political equality. They emphasized that quadratic voting, under equal voting power conditions, reduces the likelihood of producing unexpected outcomes compared to weighted voting. However, some were skeptical about whether the decided outcomes would be implemented in LLM models, suggesting guidelines at the government level to ensure compliance.

Related Work

Video Analysis in Practice & Multimodal Generative Vision Models.

Videos are a rich source of information for communication (Chen and Jiang 2019; Lin et al. 2021), driving tasks like video captioning, question answering (Yang et al. 2021; Tseng et al. 2025; Sharma et al. 2023b), text-video retrieval (Gabeur et al. 2020; Bain et al. 2021; Anne Hendricks et al. 2017). Identifying key visual content in video-language learning remains a challenge (Buch et al. 2022; Lei, Berg, and Bansal 2022). Political science research increasingly explores video content (Hong et al. 2021) where language models often exhibit biases in multi-modal data. Advancements in computer vision have led to foundational vision-language models, such as CLIP in numerous downstream applications, ranging from object detection to 3D applications (Bangalath et al. 2022; Liang et al. 2023; Rozenberszki, Litany, and Dai 2022; Ni et al. 2022), and adapted for video applications (Ni et al. 2022; Wang, Xing, and Liu 2021; Rasheed et al. 2023). More recently, multimodal integration has advanced with models like Flamingo (Alayrac et al. 2022), BLIP-2 (Li et al. 2023a) MiniGPT-4 (Zhu et al. 2023), and LLaVA (Liu et al. 2024) leveraging web-scale image-text data for improved multimodal chat capabilities. Some works extend LLMs for video comprehension (Maaz et al. 2023; Radford et al. 2021; Chiang et al. 2023; Li et al. 2023b; Liu et al. 2024), introducing Video-ChatGPT, model combining a video-optimizer for enhanced understanding.

AI Governance Approaches

AI governance concerns putting values or principles into practice via policies, while values define what agents (people or AI) ought to do (Shen et al. 2024). Researchers have argued that the research community broadly agrees about certain values and principles for better AI (Fjeld et al. 2020), though the question of how best to implement these principles is far from solved, and proceeding at different paces, in fits and spurts, around the world.

Regulatory Effort. Historically, much of the focus in AI governance research has been at the national and subnational levels (Calo 2017; Gasser and Almeida 2017; Scherer 2015). US efforts in AI governance currently focus on executive actions and industry collaboration. Executive Order 14011 "on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" (AI Executive Order) (The White House 2023) highlights the dual nature of AI's potential, stating that while AI can address critical challenges and enhance prosperity, productivity, innovation, and security, it also entails risks that require careful regulation. The U.S. approach to AI governance aims to ensure both safe and ethical AI development while achieving national strategic goals. In the short term, the U.S. will work with AI developers and other stakeholders to establish standards and guidelines.

Executive Order 14011 – like the more ambitious EU AI Act (eur 2021) – directs an assortment of actors to participate in the *standardization* of AI development and deployment. Standards are formal documents that guide organizations in achieving specific goals such as interoperability,

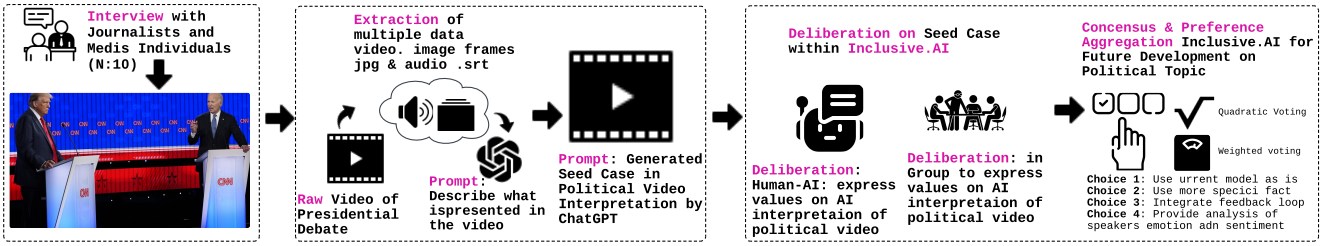

Figure 1: An overview of processes, (a) interview with experts to select suitable video example; (b) prepare seed case for experiment setup; (c) incorporate seed case into inclusive.AI system for deliberation and preference gathering (i) deliberation human-AI and group (ii) democratic voting process incorporating the voting options from deliberation of general public.

safety, or regulatory compliance. They provide a common language and practical guidelines for ensuring the safety and auditability of technical systems (Wang, Hayes, and Bashir 2022). The most rigorous standards are intended to facilitate third-party audits and certification, signaling good and trustworthy practices (Saravanakumar and Arun 2014). At present, however, the two available AI standards from NIST (AI Risk Management Framework (NIST 2023)) and ISO (ISO 42001 AI Management System (ISO 2023)) are not auditable and not yet rigorously tested.

Participatory approaches. Driven by calls from civil society organizations, academia, and others, there is growing emphasis on participatory AI governance to make AI/ML design more inclusive and equitable (Young et al. 2024; Zhang et al. 2023). However, there is limited empirical literature on involving stakeholders in refining AI performance. Frameworks like WeBuildAI (Lee et al. 2019), which facilitate participatory design for community-serving algorithms such as on demand food donation transportation services illustrate the potential of inclusive approaches. Similarly, tools like ConsiderIt (Fan and Zhang 2020), which incorporate representative deliberation in content moderation decisions, underscore the value of incorporating user voices and interests in decision-making. Crowdsourcing, as explored by Lee et al.(Lee et al. 2014), demonstrates how participatory approaches can enhance democracy by optimizing the aggregation of individual preferences into collective decisions while fostering creativity(Salganik and Levy 2015). Both in theory and practice, there exists a tension between the goal of participatory decision-making in AI and the global, distributed nature of AI development (Young et al. 2024). Kemp et al. (Erdélyi and Goldsmith 2018) and some researchers advocate for decentralized approaches, such as *"Governance Coordinating Committees,"* global standards, or leveraging existing international legal frameworks (Cihon 2019; Maas 2021; Wallach and Marchant 2018).

Yet, current AI refinement and training processes, including reinforcement learning, often rely on labor pools from developing countries due to cost considerations (Schmidt 2019). This dynamic presents a significant challenge, though not an insurmountable one, in ensuring meaningful participation on a suitable scale in AI development (Young et al. 2024). In designing AI models, it is essential to involve stakeholders and affected communities, along with AI companies, to deliberate on sensitive AI-related topics and make informed decisions about AI model behavior. However, another difficulty is establishing the best role for AI to play in group decision-making (Zheng et al. 2023).

DAO as a Tool for Governance and Co-ordination.

Decentralized Autonomous Organizations (DAOs), which emerged in the mid-2010s, share commonalities with early online communities, especially those focused on open-source projects (Chohan 2017). DAOs also draw inspiration from various models, including digital and platform cooperatives (Mannan 2018), multi-organizational networks like keiretsus (Lincoln, Gerlach, and Ahmadjian 1996), crowdfunding platforms such as Patreon, virtual economies in games like World of Warcraft and Second Life (Lehdonvirta and Castronova 2014), and peer-produced projects like Wikipedia (Xu and Li 2015). DAO governance, as a human-centric digital organization, addresses key issues in social computing but can be more complex than platforms such as civic tech (Poor 2005), and traditional online communities (Love 2010). DAOs were designed to automate organizational processes leveraging cryptographically secured blockchain technology (Buterin 2014). A key function of a DAO is collective decision-making, carried out through a series of proposals where members vote on organizational events using governance tokens, signifying relative influence within the DAO. Voting mechanisms like weighted and quadratic voting ensure secure, pseudonymous participation, with voters identified by on-chain addresses rather than real-world identities.

The emergence of DAOs introduces possible solutions, including classic coordination dilemmas such as preference aggregation, credible commitments, audience costs, information asymmetry, representation, and accountability (Hall and Taylor 1996; Wojciech Zaremba 2023). The relevance of these theories to the design of digitally-native governance institutions is a critical question (Rousseau 1964; Dahl 1989; Landemore 2012). The separation of powers in DAOs helps prevent power concentration, enhance transparency, and mitigate organizational gridlock (De Montesquieu 1989). This is increasingly relevant for AI, where inclusive decision-making is crucial throughout development lifecycle. In this work, we explore the design of DAO in AI governance for model decision-making.

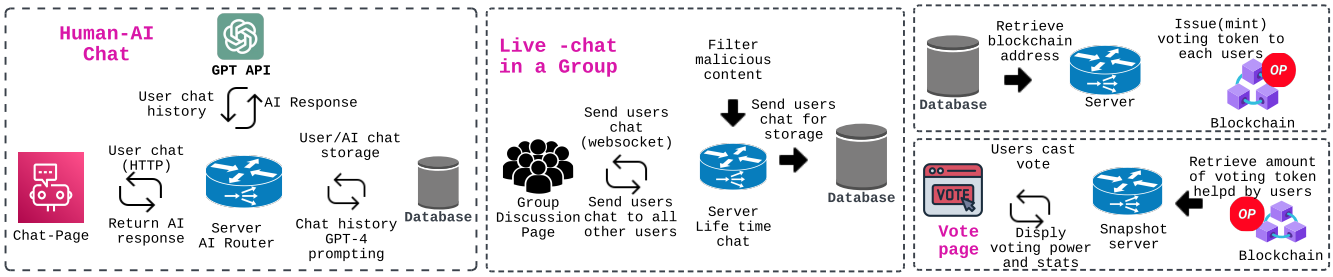

Figure 2: Incluisve.AI System Architecture

Inclusive.AI Design and Experiment

As shown in Figure 1, our entire study includes (1) an expert interview (protocol in Appendix) with journalists and media individuals in selecting a suitable political video [1] as a seed case for user experiments; (2) a large-scale user experiment in deliberating users' values regarding the LLM interpretation of political topics The user experiment has three main design components-(1) Human-AI interaction to deliberate on sensitive topics (e.g. presidential debate video), (2) Group discussion to engage with other to understand collective opinions (3) Governance decisions to guide future LLM model updates.

System Design

Inclusive.AI (GitHub (Anonymous 2024)) democratic platform (Figure 2) is deployed on the Optimism blockchain and integrates with a custom server, using Web3Auth (Goldreich 1998) for authentication. Web3Auth generates a unique Multi-Party Computation (MPC) wallet for each user, derives their blockchain address, and enables message authentication for verifying participation in votes. Upon signup, they are guided to an introduction 2-minute video overview of task details and app functionality. They then proceed to Human-AI deliberation and group discussions, where a chat box with websocket connections supports real-time interactions.

For the voting page, we implemented two VoteToken contracts using Solidity, a programming language for the Ethereum blockchain to represent users' voting power. These tokens are *minted* to users' accounts, allowing them to vote on proposals for LLM improvement of political video. The system uses the Snapshot API to create a space for governance and ensure all the processes are transparent in Blockchain. Spaces define voting rules (e.g., duration), proposal criteria (e.g., success thresholds of proposed options to be considered for LLM improvements), and roles for admins and moderators, including who can vote or propose changes. We designed spaces for each experimental condition (each type of governance decision mechanism discussed in section). When the user allocates votes accordingly and clicks the "Cast Vote," this triggers Web3Auth's signing library, which signs a message for Snapshot voting.

[1] Since we aim to understand the general public's perception of the use of LLMs for sensitive topics, such as political content, selecting politically sensitive content for the study requires careful consideration. We leveraged experts' opinions to conform to the inclusion criteria for selecting content) by providing them with an overview of the user study goal. The inclusion criteria mentioned: (a) relevance to current events (b) Broad political video (c) contextual depth or complexity (d) authenticity of content sources. We also asked them how they would prompt the LLM tool to interpret this video. We leveraged experts' feedback to design the deliberation case.

Deliberation and Decision Making

AI Guided Individual & Group Deliberation. The app begins by engaging users with an AI Value Topic related to data interpretation of a video on a political topic by GPT4. This topic is based on a 6-minute clip from the 2020 US presidential debate (Anonymous 2024) The app presented a simple question: *"Do you find the interpretation useful?" with three options (yes, no, maybe)* to stimulate further thought. Based on the user's response to the provided options, the AI continues the corresponding chat that allows users to clarify their intentions and values in natural-language conversations about AI value topics. AI resolves ambiguities through multi-turn conversations, seeking clarifications and guiding users to define their norms and expectations. Following that, users engaged in a group deliberation and learned the perspectives of others' norms. This group deliberation enables users to co-validate their values with a mini-public to make informed decisions. If participants are unable to introduce a topic on their own, they are encouraged to refer to the suggested topics provided by the tool. We designed the suggested topic based on the pilot experiment (in Appendix Section)

Democratic Decision Making for Future MM-LLM Finally, users participate in a democratic decision-making process by voting. We designed experiments to assess varying voting methods and combinations of voting power (details in section) to examine users' perception of the quality of the process being democratic in LLM model improvement decisions. We assessed users' self-reported quality with the Variety of Democracy (V-Dem) scale (Lindberg et al. 2014). The voting was live for 48 hours.

User Experimental Design

Treatment Condition: Varying Governance Voting Design In governance decision-making, voting methods and voting power are key factors influencing outcomes, as

demonstrated in DAOs and deliberative democracy (Sharma et al. 2023a, 2024b; Fritsch, Müller, and Wattenhofer 2024; Willis, Curato, and Smith 2022; Follesdal 2010). To structure decision-making to aggregate people's preferences for future LLM development, we designed a 2×2 treatment condition based on two factors: voting method and voting power, each with two levels. While alternative methods like single-choice or approval voting could also be considered, it would significantly increase the number of treatment conditions and require a large participant pool to achieve statistically significant results with actionable interpretations.

More specifically, we implemented weighted voting, commonly used in DAOs (Sharma et al. 2023a), where users distribute voting power across multiple options based on preference. To counterbalance traditional democratic aggregation which may disadvantage minority views, we incorporated quadratic voting - largely applied in real-world cases, such as Gitcoin's grant funding for public goods(Miller, Kanich, and Weyl 2024)–which enhances minority influence on crucial issues by allowing users to "pay" for additional votes. For instance, with quadratic voting, 4 tokens provide 2 votes, emphasizing the number of voters rather than voting power size (Lalley, Weyl et al. 2016). To address voting power distribution, we compared equal distribution with a Pareto-based 20/80 split, where 20% of participants receive 80% of tokens, simulating early adopters' influence. This model reflects real-world AI deployment scenarios, where certain groups benefit disproportionately.

Thus, there were four treatment conditions- (1) Quadratic Voting token-based (Participants having the same amount of token/voting power); (2) Quadratic Voting 20% population get 80% of the token as early adopters; (3) Weighted voting Token based (participants having the same amount of token/voting power); (4) Weighted voting 20% population get 80% of the token as early adopters. The goal was to assess how these variations influence users' perceptions of the process's democratic quality and outcome.

Experimental Design. Participants were randomly assigned by the Inclusive.AI system to one of four governance decision-making mechanisms, forming four treatment groups. Participants didn't know the treatment group to which they had been assigned. We employed a $2*2$ between-subjects design with 114 participants (26-30 per condition). Participants voted on four MM-LLM update options derived from 20 pilot studies for political video interpretation: (i) keep the current model; (ii) provide more specific facts; (iii) integrate a user feedback loop; (iv) analyze speakers' emotions and sentiment.

Participant Demographics

We recruited participants who are USA residents. We recruited through the CloudResearch platform (CloudResearc 2015). This study protocol involving human subjects was approved by the Institutional Review Board (IRB). Each received $30 for their participation. We used a set of screening questions. Respondents were invited to our study if they met all three selection criteria - (1) 18 years or older; (2) country of residence USA; (3) use generative AI tool. Our study resulted in total of 114 participants (Demographics in Table 2).

ID	*Gender*	*Age*	*Media Background*
E1	Female	25-34	TV News, Police issues
E2	Female	25-34	Environment, Architecture
E3	Female	25-34	Local, Under-represented
E4	Male	35-44	Political, Election
E5	Female	35-44	Weather, Political
E6	Male	35-44	TV Media
E7	Female	25-34	Economy, Tesla
E8	Male	45-54	Public Communication
E9	Male	25-34	Political
E10	Male	25-34	Local, Urban Design

Table 1: Experts demographics and background.

Experts Background.

text, images, and videos, particularly analyzing complex and sensitive topics, like US presidential debates. We recruited 10 US-based experts through personal connections and word of mouth. Experts in this study come from diverse backgrounds in journalism, media, and communication, with an equal split between males and females. Among them are Ph.D researchers specializing in media studies in urban design, and political economy, journalist focused on video media who had experience in the 2020 election coverage; medical misinformation within local communities and journalists and videographers who have covered Tesla, police issues, and local TV media, offering a unique blend of skills and perspectives

Details: Experts' Interview

Our two primary objectives of the experts' interview (Table 1) were: (i) to have a baseline of how experts envision the use of MM-LLMs for interpreting political videos to the general public and (ii) to incorporate expert feedback into the development of our methodological approach, including criteria for selecting video examples for the study. Each interview took around 1 hour.

In the first set of questions, we inquired about their primary expertise and experience with various data types, including video. This helped us understand their approach to handling different media, covering real-time events, managing diverse data for tasks like media report writing, and the factors that influence the quality of their reporting. In the next set of questions, we showed them a political video of US presidential debate and asked them to interpret it using multimodal data (e.g., audio, visuals, closed captions). We asked, *"Can you walk me through the process you employ to analyze the video content to write a report?"* Following this, we asked their opinion on using LLMs for video analysis We then showed them how LLMs (ChatGPT) interpreted the same video and asked for their thoughts to identify the benefits, limitations, and critical factors in interpreting contentious topics.

Since we aim to understand the general public's perception of the use of MM-LLMs in better designing models for sensitive topics, such as political content, selecting politically sensitive content for the study requires careful consideration. We leveraged experts' opinions to conform to the

Gender (%)			Age (%)				Race (%)				
Woman	Man	Non-binary	18-24	25-34	35-44	45-54	White	Black	Asian	Latin	Others
45.6	52.6	1.8	21.1	39.5	27.2	12.3	52.6	12.3	21.9	10.5	2.63
Education (%)											
High school	Bachelor	Masters/professional			Doctorate		College/vocational training				Others
14.0	41.2	12.3			2.6		26.3				3.5
Political Orientation (%)											
Very conservative		Conservative		Moderate			Liberal		Very liberal		
5.3		21.1		22.8			35.1		15.8		
Political Party (%)											
Republic party		Democratic party		Libertarian party			Independent/Unaffiliated				
21.9		50.9		2.6			24.6				

Table 2: Participants' demographics ($n = 114$)

inclusion criteria for selecting content (details in Appendix) by providing them with an overview of the user study goal. We also asked them how they would prompt the LLM tool to interpret this video. We leveraged experts' feedback to design the deliberation seed usecase.

Data Analysis

For qualitative data, two researchers independently read through 20% of the individual deliberation with AI-agent and group discussion data, developed codes, and compared them until we developed a consistent codebook. We met regularly to discuss the coding and agreed on a shared codebook. Once the codebook was finalized, two researchers divided the remaining data and coded them. After completing coding for all individual and group deliberation, both researchers spot-checked other's coded transcripts and did not find any inconsistencies. Finally, we organized our codes into higher-level categories. We followed a deductive thematic analysis to explore participants' values, expectations towards MM-LLM video interpretation (Clarke and Braun 2017). We grouped lower-level codes into sub-themes and further extracted main themes. We used a similar approach for the qualitative analysis of expert interview data to identify high-level themes. For the quantitative analysis of users' perceptions of DAO governance mechanisms, the data analysis process is explained directly within the respective results.

People's Opinion on LLM Interpretation of Political Video

Journalist's Opinion of LLM in Political Video.

We found several practices of journalists in interpreting political videos on their own, including- (a) fact-checking with multiple data sources and guidelines (e.g. media literacy project, MSA Security) (b) involvement of expert-in-the-loop (e.g. academic scholars, senior journalists, domain experts), (c) narrative approach considered as news generation 101; (d) theoretical underpinning, such as positionality, selective exposure (Tully et al. 2022).

Experts highlighted several limitations in LLM-generated summaries of political videos, particularly the absence of human interaction cues such as tone and emotion. They noted that the lack of contextual information, including background knowledge on political debates, reduced the summary's usefulness for news content. While factually accurate, the summary failed to capture the antagonistic and dramatic dynamics of the debate, including conflicts, personal attacks, and the candidates' lack of factual references. Additionally, experts criticized its lack of storytelling and engagement, making it unsuitable for a diverse audience and insufficient in depth and impact.

General Public's Opinion.

Our findings of users' interaction with the seed case on political video interpretation highlight various factors participants considered important on interpreting video content while analyzing multiple types of data (e.g. image frame, audio, etc). In group deliberation, we found that participants articulated their arguments in longer sentences, while in human-AI chats, the conversations were shorter. In individual value elicitation, we also found participants to suggest specific design recommendations of how to generate and present the LLM output rather than only pointing out what is lacking. They tend to begin their interactions with a positive tone. As the conversations progressed, participants shifted towards making recommendations and expressing concerns. In contrast, group deliberation started with a tone of concern and debate.

However, in both types of interactions, there are overlapping values emerged regarding LLM improvement for political content interpretation. This includes: the emotions of the speaker, subjective content (e.g., who supports or opposes, composure, professionalism), and the speaker's positionality. We also observed nuanced differences in individual values, for instance, participants tend to express a preference for fact-focused political LLM interpretation with specific indicators as design recommendations and emphasized the importance of clarity and organization of LLM output (Table 3 presents example quotes).

Experience in Democratic Governance

Preference on LLM Improvement Choices.

For improving MM-LLMs in political video interpretation, participants strongly preferred *"providing more spe-*

Theme	Quote	Ind / Group
Emotion of Speakers	*"There was a heated argument in video, both speakers didn't want to give way for other to speak, Trump and moderator were talking like they were fighting, its not in the LLM output."*	Indv & Group
Objectivity of Situation	*"It didn't understand situation at all, AI was superficial, capturing the scene, distinguished between speaker and their political view is important, I could have just read the subtitle instead."*	Indv & Group
Desire FactChecking	*"Fact-checking whether the debaters are saying anything of substance would greatly help in giving an accurate picture of the view, like citing some source while interpreting the video."*	Indv & Group
Balance Brevity and Substance	"AI describe he video, but there was no real context, like to take away"	Group
Balancing Content & Biases	"This is one of those times when I wish AI could let itself loose just a little more, necessarily—just to the fact of acknowledging how Trump was not acting as a good steward of discussion, also the too much emphasize towards Obamacare and social support system."	Group
Organization of LLM output	"It would be easier to differentiate to have the description of two candidates side by side."	Indv
Specific Design Recommendations,	"Red highlight for content in the video that are factually wrong and green for truth."	Indv

Table 3: Overview of Themes of Deliberation on LLM Output of Political Video

cific facts"(choice 2), followed by *"analyzing speakers' emotions and sentiment"* (choice 4) and *"integrating a user feedback loop"* (choice 3). The consistency of choices 2 and 4 across quadratic and weighted methods indicates stable user preference (Table **??**). However, in 20/80 voting power distributions, early adopters (80% power) influenced outcomes, narrowing the gap between choices 3 and 4. This suggests that in real-world governance of LLM improvements, decision-making that concentrates power among a few influential stakeholders could disproportionately shape LLM improvements, potentially misaligning with broader user preferences.

To see whether participants affiliated with different political parties had different choices and perceptions for LLM improvement on political video interpretation, we ran a linear regression controlling for voting methods. As a result, we found that, compared to Democrats, Republicans were less likely to vote for Choice 2 (i.e., provide more specific facts) with a P-value of 0.084 and more likely to vote for Choice 3 (i.e., integrate a user feedback loop) with a P-value of 0.054.

General Perception of Voting Mechanism.

With a 5-point Likert scale (Figure 3), we found participants' perceptions of the voting process usage in LLM governance where most participants were satisfied with the process regardless of voting protocols. Notably, they rated with average scores of $3.89, 4.17, 3.96$, and 3.93 in the four voting mechanisms: quadratic+equal, quadratic+20/80, ranked+equal, and ranked+20/80. Quadratic voting and equal power distribution enhanced participants' trust in the decision-making process, reducing concerns about unexpected outcomes. As participants shared *"I split my votes across multiple issues, but I think this is the purpose—to vote carefully for the option I care about most. It allows stronger opinions on some issues. the square thing I like, so even if sometimes someone had more token than me, that's actually not the number that would apply rather square root."* A linear regression analysis confirmed this effect: the coefficient for quadratic $0.4772 (P = 0.013)$, for same was $0.4002 (P = 0.038)$. The linear regression considering the

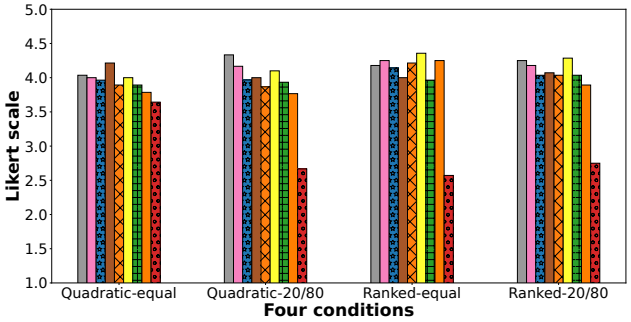

Figure 3: Users' perception of the quality of voting mechanism in governance decision making. Each bar is for the following item: 1) I found the voting method meaningful to include my voice, 2) I found this voting method relevant with the purpose of the proposal, 3) I found that voting power meaningful in including my voice, 4) I felt that I can contribute shaping the Generative AI model, 5) I found this voting power relevant with the purpose of the proposal, 6) I found this voting method fair, 7) I felt I have some power to affect Generative AI models for future development, 8) I found voting power distribution among users equitable, and 9) I felt the voting power distribution would not result in unexpected outcome

	Choice 1		Choice 2		Choice 3		Choice 4	
	mean	std.	mean	std	mean	std	mean	std.
Quadratic - same (n: 29)	0.0814	0.0885	0.4300	0.3421	0.1524	0.1690	0.2597	0.2905
Quadratic -20/80 (n: 30)	0.1193	0.2197	0.3267	0.2260	0.2188	0.1956	0.3080	0.2473
Ranked - same (n: 27)	0.1193	0.1412	0.3941	0.2255	0.1896	0.1197	0.2926	0.2351
Ranked - 20/80 (n:28)	0.1395	0.1534	0.4044	0.2692	0.1877	0.1400	0.2417	0.1870

Table 4: Summary stats of the ratio of tokens allocated to each voting choice (Choice 1: Keep the current model, Choice 2: Provide more specific facts, Choice 3: Integrate a user feedback loop, and Choice 4: Analyze speakers' emotions and sentiment) by users. The ratio is calculated as the percentage of tokens the user allocated to each voting option. For example, if a user allocated 20, 20, 30, 30 tokens for each voting option, the vector for the user would be (0.2,0.2,0.3,0.3).

Democrats	**Republicans**	**Independent / Unaffiliated**
Progressive and Empowerment	Flexible in Distributing power	Desire for Additional Option
Ease of Use and Intuitive	New Experience and Curiosity	Ease of Use and Intuitive
Support Multiple Choices	Quantifying Perception and Thinking Critically	Weighted Voting as a Preferred Feature
Quadratic Voting Perceived as Fair	Having Influence on AI Development	Applying This Process to Other Contexts
Engaging and Enjoyable	Concerns About Complexity and Restrictions	Concerns of Fairness and Transparency
Informed and accurate decision-making	Concerns About External Influences and Bias	Concerns About the Process's Impact

Table 5: Governance Decision-Making Experience Across Different Political Leaning

interaction also demonstrated statistical significance; the coefficient of *quadratic*×*same* was 1.1548 with a P-value of 0.002.

Quality of Decision-Making Process of Different Democracies.

We examined participants' perceptions of LLM governance using the Varieties of Democracy (V-Dem) (Figure 4). xf. As noted, *"The voting was inclusive—I would like this process in chatGPT like system where they broadcast such voting time to time to get some signal from users rather deploying by themselves only."* This supports the argument that active user participation in AI decision-making can enhance legitimacy, rather than centralized deployment by developers. Voting power distribution further reinforced perceptions of political equality (coefficient= 0.8091, P-value< 0.001), with linear regression considering interaction confirming its significance (coefficient of *same*= 0.7500, P-value= 0.019). This highlights the need of fair representation in AI oversight, where users regardless of their expertise or influence should have a say in shaping AI behavior.

Relationship Between Users' Value Towards AI and Their Perceived Democracy Value

Participants who found LLM personally relevant were more likely to view the DAO-enabled voting process as highly participatory (Pearson Corr= −0.4426, P-value< 0.001). This underscores the need for AI systems to establish personal relevance with users, potentially through more user-centered political content moderation. Perceptions of deliberative democracy were strongly linked to trust in AI companies(Pearson Corr= 0.4422, P-value< 0.001) and perceived AI risks (Pearson Corr= 0.5142, P-value< 0.001). This suggests that skepticism about AI risks coexists with the belief that AI governance should involve ongoing public discourse. For LLM governance, this emphasizes the need for mechanisms that allow users to contest, audit, and deliberate on AI-generated political content, rather than simply consuming it. Participants who valued civil liberties also emphasized the importance of diverse datasets in AI training (Pearson Corr= 0.4646, P-value< 0.001), uncertainty han-

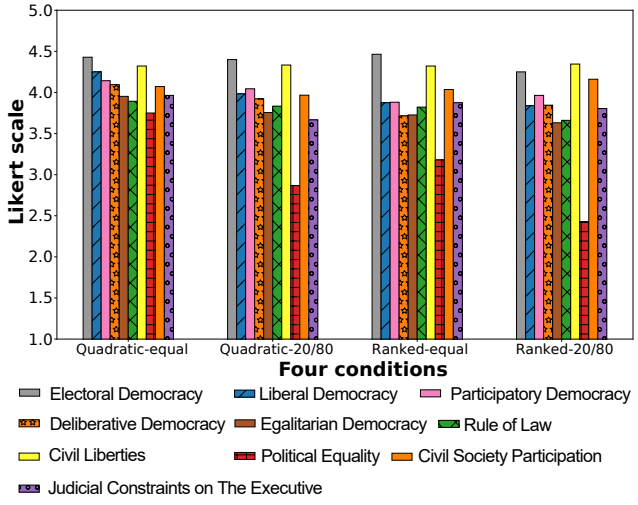

Figure 4: Users' perception of a voting mechanism (obtained through the V-Dem question lists)

dling by AI developers (Pearson Corr$= 0.5326$, P-value< 0.001), the perceived AI risks (Pearson Corr$= 0.4407$, P-value< 0.001), and the desired reliance on AI (Pearson Corr$= 0.4950$, P-value< 0.001). This underscores the necessity of dataset diversity, bias mitigation, and AI uncertainty management in political content generation.

Attitude Towards Voting Mechanisms.

Participants had key attitude in applying voting mechanisms to MM-LLM governance, including (1) progressive and fair process, (2) methods to show the strength of preference, (3) support multiple choices with a unique voice, (4) quantifying perception, (5) Inclusive.AI as practical AI governance applications (e.g., aligning with public preferences). We also found differences in the perception among political parties. Republicans tended to feel significantly more that they could contribute to shaping the space of generative AI models through this process when compared to Democrats (linear coefficient$= 0.521$, P-value$= 0.007$). Qualitative analysis of survey data also revealed some differing perspectives (Figure **??**). Democrats emphasized empowerment, ease of use, and engagement emphasizing positive experience. In contrast, republicans prioritized functional and individual priorities like flexible voting power designs, on option to quantify perception through voting, and curiosity. Republican and independent participants also raised concerns about complexity, external influences (majority bias as a good way to go), and post-vote transparency regarding AI developers' implementation of decisions. However, these findings are not indicative of broader political divisions due to the low frequency of such experiences.

Discussion

Our findings underscore two key recommendations for practitioners aligning with LLMs and how to engage users in governance.

DAO as a Technical LLM Governance Solution.

Transparency in LLM design decisions is utmost importance for aligning AI systems with societal expectations (Mitchell et al. 2019; Liesenfeld, Lopez, and Dingemanse 2023). To do that, it's crucial not only to gain a deeper understanding of public perceptions regarding AI but also to devise methods that actively involve the community in the decision-making processes governing AI technologies. Inclusive.AI tool, underpinned by the DAO mechanism, offers an avenue to actively involve people in governing llm with empirical evidence while presented with a sensitive topic like political video. DAO mechanisms, as digital-first entities, employ mechanisms like initiating proposals, nuanced voting methods, and blockchain-based coordination (Sharma et al. 2023a), offering a structured approach to AI governance (Koster et al. 2022), a concept endorsed by industry leaders such as OpenAI, Meta, and federal agencies (Wojciech Zaremba 2023; Biden 2023).

A standout feature was our system's Voting method, in which participants found effectively representing their voice directly impacting AI model decisions for future improvements (Arts and Tatenhove 2004). Participants recognized the potential of these methods in helping developers and government bodies align more closely with public preferences. However, skepticism remains about whether their votes would translate into real changes in AI models. This highlights the need for government-level guidelines to ensure system compliance and evidence through future audits.

Continuous Human Involvement for LLM Model Adaptability.

Our research, drawing on insights from experts in news production reveals that video analysis in media coverage remains largely reliant on manual processes and human intervention. Critical frameworks like positionality (Callison and Young 2019) and selective exposure balance are essential for ensuring accurate and contextually rich video interpretation, particularly in political reporting. Experts emphasize the need for diverse perspectives and contextual depth to prevent biases and ensure political content reflects a broad spectrum of viewpoints (Blumler and Kavanagh 1999; Jacobs and Townsley 2011)

Our findings from public deliberation and LLM governance decision-making illustrate how political affiliation shapes perceptions of LLM-generated political content. For instance, Republicans were less likely than Democrats to vote for providing more specific facts and instead favored integrating a user feedback loop for LLM improvement. InclusiveAI platform facilitate on imitating natural human interaction among people acknowledging that conflicting interests and preferences. Rather than seeking consensus on the topic, participants engaged in discussions that helped them identify compromises and make informed voting decisions. We suggest that inclusive AI systems could be integrated into LLM tools, such as ChatGPT, allowing users and experts to propose real-time adjustments and engage with broader communities when necessary. It highlights a potential future where people continuously engage in shaping functionality of AI systems with evolving needs. Potential risks of this research include political bias reinforcement due to differing perceptions of LLM-generated content, potentially deepening ideological divides.

Future Deployment and Evaluation of Inclusive.AI

Stakeholders involved in the governance decision-making in LLM improvement felt their contributions were meaningful, particularly in the configurations of quadratic voting with equal power distribution facilitates minority voices. These results indicate that a DAO-enabled governance process can increase stakeholder satisfaction in AI decision-making compared to centralized approaches, where only a very small group of stakeholders (e.g., frontier AI companies) make the decision. However, we acknowledge that deployment of our system in the wild can further broaden the impact. One concrete way that we propose to incorporate Incluisve.AI tool with LLM applications (e.g., ChatGPT, DALL·E, Perplexity, etc) or LLM-powered application (e.g, health chatbot, legal interpretation chatbot) is that Inclusive.AI could be integrated as a plugin within ChatGPT that enables users to challenge and collaboratively deliberate over AI-generated content whether it is a summary of

a political debate, a medical explanation, an image interpretation, or a cultural analysis. When users find an output unsatisfactory, biased, or harmful, they can press a *"Challenge this Response"* button, opening a structured interface to provide alternative views, engage in a real-time deliberation thread, and vote on LLM improvement suggestions. The system can capture people's varying perceptions and flags contested outputs for model refinement. Over time, it can aggregate user preferences across domains, politics, health, education, and social issues to inform transparent and participatory AI behavior tuning.

To extend Inclusive.AI beyond this study, we envision a set of realistic studies modeling contentious AI outputs that might provoke public disagreement. In one scenario, an AI system generates misleading vaccine safety information (Ahmad et al. 2025; Xu et al. 2025)— Inclusive.AI can enable deliberation among patients, public health experts, and ethicists, evaluating outcomes such as perceived legitimacy while media outlets can integrate it to audit neutrality in political captioning (Edenberg and Wood 2023; Norris 2009); and civic platforms may apply it to generative summaries of legislation (Fan and Zhang 2020; Tsai et al. 2024). In the future, we aim to evaluate Inclusive.AI's deliberative refinement model against top-down moderation (Seering 2020), non-deliberative crowd voting (Fishkin and Luskin 2003; Gerber et al. 2014; Sharma et al. 2022), and expert-only audits (Zhang and Dafoe 2020; Costanza-Chock, Raji, and Buolamwini 2022) to assess trust, transparency, and correction efficacy.

Ethical Considerations in Democratic AI Governance

Inclusive.AI in practice requires careful attention to ethical and legal factors to ensure that the governance process itself upholds inclusivity, fairness, and complies with regulations while people's identity and interaction are involved. Our system addresses these challenges by minimizing personal data on-chain and using pseudonymized identifiers for participants to engage in deliberation and democratic voting in LLM improvement. Depending on the use case, one can enable a setting to keep sensitive user information (e.g., demographic details or discussion content) off-chain, while decision records remain immutable and transparent. Additionally, to further ensure privacy, current infrastructure can adapt with emerging solutions like confidential DAOs which encrypt governance data yet maintain transparent, immutable logs (zam 2025; Sharma et al. 2024a) as well as maintain confidence among stakeholders that participation will not expose their personal data and will comply with privacy regulations (e.g., GDPR (Haque et al. 2021; Potter et al. 2025)) and ethical norms.

Another critical consideration is preventing the concentration of power and ensuring fairness within the DAO itself. A known pitfall in some blockchain governance systems is the risk of "whale" voters or token-based plutocracy, where a small number of actors accumulate disproportionate influence (Sharma et al. 2023a; Halaburda 2025). We mitigate this risk through system design, such as, quadratic voting to curb dominance by any single group. For example, quadratic voting is used to give diminishing returns on additional votes, so that participants must spend votes quadratically to express stronger preferences (Lalley and Weyl 2018). This mechanism ensures that minority opinions can still influence outcomes on issues they care deeply about, counterbalancing majority rule. In practice, quadratic voting allowed marginalized voices in our study to impact AI design decisions significantly more than under a traditional one-person-one-vote system. In the future, we will explore alternatives to purely token-based governance or a hybrid approach, for instance, sortition-based DAOs which enable a rotating panel of randomly chosen participants, as a *"governance jury,"* for unbiased participant selection, ensuring no clique can continuously dominate decisions (Kelsey 2023). These combined measures are compatible with the current infrastructure of Inclusive.AI to prevent power concentration with fair allocation of voice, advanced voting schemes, and hybrid governance models.

Finally, we designed the system to embed values with traceability throughout the AI governance lifecycle. It addresses the "black box" concern often raised in AI ethics since decisions made via Incluisve.AI are traceable in blockchain. Stakeholders and auditors can inspect how a particular AI model update was agreed upon, which arguments were presented. This not only builds trust but also facilitates external oversight or audits for compliance. In case of regulatory review or legal disputes, the immutable governance record serves as evidence of due diligence and community consent in the AI's decision-making process.

Conclusion

Our study demonstrates that decentralized, participatory governance mechanisms in Inclusive.AI platform can support large language model (LLM) development decision making with diverse public values, particularly in politically sensitive contexts. By combining individual and collective deliberation with structured voting methods like quadratic and weighted voting, the platform enabled meaningful engagement from users across political affiliations. Our research suggests that integrating democratic input into AI governance not only enhances user satisfaction but also offers a scalable model for aligning AI behavior with societal expectations.

Ethics Statement

This study protocol involving human subjects was approved by the Institutional Review Board (IRB). The data collection and transcription generation was anonymous to preserve privacy of the users. This study explores decentralized governance mechanisms in decision-making for LLM improvement by engaging users, particularly in politically sensitive contexts. The InclusiveAI tool with transparent design, equitable participation can allow to shape AI with broader perspectives. This also has a future potential to potentially involve regulatory oversight for for the responsible implementation.

Positionality Statement

All authors are currently affiliated with US academic institutions, and all our interviewees also live in the United States. Although none of the team members have extensive experience in politics or journalism, the research team includes members who have long been engaged in AI/ML, Human Centered Computing and privacy and security research, as well as members with extensive experience in communication and decentralized applications research experience, which ensures that we can understand the problems faced in AI governance from multiple perspectives.

Adverse Impact Statements

As the paper only contains non-identifiable interview, survey, deliberations and voting tally. All of the data is collected with anonymity, which is one of the fundamental functionalities of the decentralized system infrastructure implemented in Inclusive.AI. We do not expect that the dissemination of this paper will have a substantial adverse impact.

Acknowledgements

We thank OpenAI for supporting this research through the "Democratic Input to AI Grant 2023." We also would like to thank Teddy Lee, and Tyna Eloundou for their feedback.

References

2021. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS.

2025. Confidential DAO and governance systems using Fully Homomorphic Encryption - Zama — zama.ai. https://www.zama.ai/solutions/confidential-dao-using-fully-homomorphic-encryption. [Accessed 08-05-2025].

Ahmad, S. T.; Lu, H.; Liu, S.; Lau, A.; Beheshti, A.; Dras, M.; and Naseem, U. 2025. VaxGuard: A Multi-Generator, Multi-Type, and Multi-Role Dataset for Detecting LLM-Generated Vaccine Misinformation. *arXiv preprint arXiv:2503.09103*.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Anne Hendricks, L.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, 5803–5812.

Anonymous. 2024. GitHub. https://github.com/AccountProject/Inclusive.AI-MM_LLM. [Accessed 14-02-2025].

Arts, B.; and Tatenhove, J. V. 2004. Policy and power: A conceptual framework between the 'old'and 'new'policy idioms. *Policy sciences*, 37: 339–356.

Bain, M.; Nagrani, A.; Varol, G.; and Zisserman, A. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1728–1738.

Bangalath, H.; Maaz, M.; Khattak, M. U.; Khan, S. H.; and Shahbaz Khan, F. 2022. Bridging the gap between object and image-level representations for open-vocabulary detection. *Advances in Neural Information Processing Systems*, 35: 33781–33794.

Barocas, S.; Hood, S.; and Ziewitz, M. 2013. Governing algorithms: A provocation piece. *Available at SSRN 2245322*.

Benkler, Y.; Shaw, A.; and Hill, B. M. 2015. Peer production: A form of collective intelligence. *Handbook of collective intelligence*, 175.

Biden, J. R. 2023. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence.

Blumler, J. G.; and Kavanagh, D. 1999. The third age of political communication: Influences and features. *Political communication*, 16(3): 209–230.

Brundage, M.; Avin, S.; Wang, J.; Belfield, H.; Krueger, G.; Hadfield, G.; Khlaaf, H.; Yang, J.; Toner, H.; Fong, R.; et al. 2020. Toward trustworthy AI development: mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213*.

Bu, F.; Wang, N.; Jiang, B.; and Liang, H. 2020. "Privacy by Design" implementation: Information system engineers' perspective. *International Journal of Information Management*, 53: 102124.

Buch, S.; Eyzaguirre, C.; Gaidon, A.; Wu, J.; Fei-Fei, L.; and Niebles, J. C. 2022. Revisiting the" video" in video-language understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2917–2927.

Butcher, J.; and Beridze, I. 2019. What is the state of artificial intelligence governance globally? *The RUSI Journal*, 164(5-6): 88–96.

Buterin, V. 2014. DAOs, DACs, DAs and more: An incomplete terminology guide. *Ethereum Blog*, 6: 2014.

Callison, C.; and Young, M. L. 2019. *Reckoning: Journalism's limits and possibilities*. Oxford University Press.

Calo, R. 2017. Artificial intelligence policy: a primer and roadmap. *UCDL Rev.*, 51: 399.

Chen, S.; and Jiang, Y.-G. 2019. Motion guided spatial attention for video captioning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 8191–8198.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3): 6.

Chohan, U. W. 2017. The decentralized autonomous organization and governance issues. *Available at SSRN 3082055*.

Cihon, P. 2019. Standards for AI governance: international standards to enable global coordination in AI research & development. *Future of Humanity Institute. University of Oxford*, 340–342.

Clarke, V.; and Braun, V. 2017. Thematic analysis. *The journal of positive psychology*, 12(3): 297–298.

CloudResearc. 2015. CloudResearch. https://www.cloudresearch.com/. [Accessed 15-02-2025].

Costanza-Chock, S.; Raji, I. D.; and Buolamwini, J. 2022. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1571–1583.

Costello, T. H.; Pennycook, G.; and Rand, D. G. 2024. Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714): eadq1814.

Dahl, R. 1989. Democracy and its critics Yale university press. *New Haven & London*.

De Montesquieu, C. 1989. *Montesquieu: The spirit of the laws*. Cambridge University Press.

Edenberg, E.; and Wood, A. 2023. Disambiguating algorithmic bias: From neutrality to justice. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 691–704.

Erdélyi, O. J.; and Goldsmith, J. 2018. Regulating artificial intelligence: Proposal for a global solution. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 95–101.

Fan, J.; and Zhang, A. X. 2020. Digital juries: A civics-oriented approach to platform governance. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–14.

Feng, S.; Park, C. Y.; Liu, Y.; and Tsvetkov, Y. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. *arXiv preprint arXiv:2305.08283*.

Fisher, J.; Feng, S.; Aron, R.; Richardson, T.; Choi, Y.; Fisher, D. W.; Pan, J.; Tsvetkov, Y.; and Reinecke, K. 2024. Biased AI can Influence Political Decision-Making. *arXiv preprint arXiv:2410.06415*.

Fishkin, J. S.; and Luskin, R. C. 2003. The quest for deliberative democracy. In *Democratic innovation*, 31–42. Routledge.

Fjeld, J.; Achten, N.; Hilligoss, H.; Nagy, A.; and Srikumar, M. 2020. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI.

Follesdal, A. 2010. The place of self-interest and the role of power in the deliberative democracy. *Journal of political philosophy*, 18(1): 64–100.

Fritsch, R.; Müller, M.; and Wattenhofer, R. 2024. Analyzing voting power in decentralized governance: Who controls DAOs? *Blockchain: Research and Applications*, 100208.

Gabeur, V.; Sun, C.; Alahari, K.; and Schmid, C. 2020. Multi-modal transformer for video retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, 214–229. Springer.

Gasser, U.; and Almeida, V. A. 2017. A layered model for AI governance. *IEEE Internet Computing*, 21(6): 58–62.

Gerber, M.; Bächtiger, A.; Fiket, I.; Steenbergen, M.; and Steiner, J. 2014. Deliberative and non-deliberative persuasion: Mechanisms of opinion formation in EuroPolis. *European Union Politics*, 15(3): 410–429.

Goldreich, O. 1998. Secure multi-party computation. *Manuscript. Preliminary version*, 78(110): 1–108.

Halaburda, H. 2025. The hidden danger of re-centralization in blockchain platforms. https://www.brookings.edu/articles/the-hidden-danger-of-re-centralization-in-blockchain-platforms. [Accessed 08-05-2025].

Hall, P. A.; and Taylor, R. C. 1996. Political science and the three new institutionalisms. *Political studies*, 44(5): 936–957.

Haque, A. B.; Islam, A. N.; Hyrynsalmi, S.; Naqvi, B.; and Smolander, K. 2021. GDPR compliant blockchains–a systematic literature review. *Ieee Access*, 9: 50593–50606.

Hong, J.; Crichton, W.; Zhang, H.; Fu, D. Y.; Ritchie, J.; Barenholtz, J.; Hannel, B.; Yao, X.; Murray, M.; Moriba, G.; et al. 2021. Analysis of faces in a decade of us cable tv news. In *KDD'21: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*.

ISO. 2023. ISO/IEC 42001.

Jacobs, R. N.; and Townsley, E. 2011. *The space of opinion: Media intellectuals and the public sphere*. Oxford University Press.

Ji, J.; Liu, M.; Dai, J.; Pan, X.; Zhang, C.; Bian, C.; Chen, B.; Sun, R.; Wang, Y.; and Yang, Y. 2024. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36.

Kaminski, M. E.; and Malgieri, G. 2020. Multi-layered explanations from algorithmic impact assessments in the GDPR. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 68–79.

Kelsey, J. 2023. Implementing a Sortition-based DAO for Policymaking and AI Governance. *Sat*, 5: 05.

Kemper, J.; and Kolkman, D. 2019. Transparent to whom? No algorithmic accountability without a critical audience. *Information, Communication & Society*, 22(14): 2081–2096.

Koster, R.; Balaguer, J.; Tacchetti, A.; Weinstein, A.; Zhu, T.; Hauser, O.; Williams, D.; Campbell-Gillingham, L.; Thacker, P.; Botvinick, M.; and Summerfield, C. 2022. Human-Centred Mechanism Design with Democratic AI. 6(10): 1398–1407.

Lalley, S. P.; and Weyl, E. G. 2018. Quadratic voting: How mechanism design can radicalize democracy. In *AEA Papers and Proceedings*, volume 108, 33–37. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Lalley, S. P.; Weyl, E. G.; et al. 2016. Quadratic voting. *Available at SSRN*.

Landemore, H. 2012. *Democratic reason: Politics, collective intelligence, and the rule of the many*. Princeton University Press.

Lee, D.; Goel, A.; Aitamurto, T.; and Landemore, H. 2014. Crowdsourcing for participatory democracies: Efficient elicitation of social choice functions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 2, 133–142.

Lee, M. K.; Kusbit, D.; Kahng, A.; Kim, J. T.; Yuan, X.; Chan, A.; See, D.; Noothigattu, R.; Lee, S.; Psomas, A.; et al. 2019. WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on human-computer interaction*, 3(CSCW): 1–35.

Lehdonvirta, V.; and Castronova, E. 2014. *Virtual economies: Design and analysis*. Mit Press.

Lei, J.; Berg, T. L.; and Bansal, M. 2022. Revealing single frame bias for video-and-language learning. *arXiv preprint arXiv:2206.03428*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, K.; He, Y.; Wang, Y.; Li, Y.; Wang, W.; Luo, P.; Wang, Y.; Wang, L.; and Qiao, Y. 2023b. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; and Marculescu, D. 2023. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7061–7070.

Liesenfeld, A.; Lopez, A.; and Dingemanse, M. 2023. Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th international conference on conversational user interfaces*, 1–6.

Lin, X.; Bertasius, G.; Wang, J.; Chang, S.-F.; Parikh, D.; and Torresani, L. 2021. Vx2text: End-to-end learning of video-based text generation from multimodal inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7005–7015.

Lincoln, J. R.; Gerlach, M. L.; and Ahmadjian, C. L. 1996. Keiretsu networks and corporate performance in Japan. *American sociological review*, 67–88.

Lindberg, S. I.; Coppedge, M.; Gerring, J.; and Teorell, J. 2014. V-Dem: A new way to measure democracy. *Journal of Democracy*, 25(3): 159–169.

Linegar, M.; Kocielnik, R.; and Alvarez, R. M. 2023. Large language models and political science. *Frontiers in Political Science*, 5: 1257092.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Love, R. 2010. *Linux kernel development*. Pearson Education.

Maas, M. M. 2021. Aligning AI regulation to sociotechnical change. *Oxford Handbook on AI Governance (Oxford University Press, 2022 forthcoming)*.

Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. S. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

Mannan, M. 2018. Fostering worker cooperatives with blockchain technology: Lessons from the Colony project. *Erasmus L. Rev.*, 11: 190.

Miller, J.; Kanich, C.; and Weyl, E. G. 2024. A Case Study in Plural Governance Design. In *appear at the Pluralistic Alignment Workshop at NeurIPS*.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, 220–229.

Ni, B.; Peng, H.; Chen, M.; Zhang, S.; Meng, G.; Fu, J.; Xiang, S.; and Ling, H. 2022. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*, 1–18. Springer.

NIST. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0).

Norhashim, H.; and Hahn, J. 2024. Measuring Human-AI Value Alignment in Large Language Models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 1063–1073.

Norris, P. 2009. *Public Sentinel: News media and governance reform*. World Bank Publications.

Poor, N. 2005. Mechanisms of an online public sphere: The website Slashdot. *Journal of computer-mediated communication*, 10(2): JCMC1028.

Potter, Y.; Choi, Y.; Rand, D.; and Song, D. 2024a. LLMs' Potential Influences on Our Democracy: Challenges and Opportunities. In *ICLR Blogposts 2025*. Accessed: 2025-01-02.

Potter, Y.; Corren, E.; Garrido, G. M.; Hoofnagle, C.; and Song, D. 2025. The Gap Between Data Rights Ideals and Reality. *arxiv*.

Potter, Y.; Lai, S.; Kim, J.; Evans, J.; and Song, D. 2024b. Hidden Persuaders: LLMs' Political Leaning and Their Influence on Voters. *arXiv preprint arXiv:2410.24190*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rasheed, H.; Khattak, M. U.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6545–6554.

Rousseau, J.-J. 1964. The social contract (1762). *Londres*.

Rozado, D. 2024. The political preferences of LLMs. *arXiv preprint arXiv:2402.01789*.

Rozenberszki, D.; Litany, O.; and Dai, A. 2022. Language-grounded indoor 3d semantic segmentation in the wild. In *European Conference on Computer Vision*, 125–141. Springer.

Rubinstein, I. S.; and Good, N. 2013. Privacy by design: A counterfactual analysis of Google and Facebook privacy incidents. *Berkeley Tech. LJ*, 28: 1333.

Salganik, M. J.; and Levy, K. E. 2015. Wiki surveys: Open and quantifiable social data collection. *PloS one*, 10(5): e0123483.

Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, 29971–30004. PMLR.

Saravanakumar, C.; and Arun, C. 2014. Survey on interoperability, security, trust, privacy standardization of cloud computing. In *2014 International Conference on Contemporary Computing and Informatics (IC3I)*, 977–982. IEEE.

Scherer, M. U. 2015. Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harv. JL & Tech.*, 29: 353.

Schmidt, F. A. 2019. Crowdsourced production of AI training data: how human workers teach self-driving cars how to see. Technical report, Working Paper Forschungsförderung.

Seering, J. 2020. Reconsidering self-moderation: the role of research in supporting community-based models for online content moderation. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2): 1–28.

Selbst, A. D. 2021. An institutional view of algorithmic impact assessments. *Harv. JL & Tech.*, 35: 117.

Sharma, T.; Kwon, Y.; Pongmala, K.; Wang, H.; Miller, A.; Song, D.; and Wang, Y. 2023a. Unpacking How Decentralized Autonomous Organizations (DAOs) Work in Practice. *arXiv preprint arXiv:2304.09822*.

Sharma, T.; Nair, V. C.; Wang, H.; Wang, Y.; and Song, D. 2024a. "I Can't Believe It's Not Custodial!": Usable Trustless Decentralized Key Management. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–16.

Sharma, T.; Potter, Y.; Pongmala, K.; Wang, H.; Miller, A.; Song, D.; and Wang, Y. 2024b. Future of Algorithmic Organization: Large-Scale Analysis of Decentralized Autonomous Organizations (DAOs). *arXiv preprint arXiv:2410.13095*.

Sharma, T.; Stangl, A.; Zhang, L.; Tseng, Y.-Y.; Xu, I.; Findlater, L.; Gurari, D.; and Wang, Y. 2023b. Disability-First Design and Creation of A Dataset Showing Private Visual Information Collected With People Who Are Blind. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–15.

Sharma, T.; Zhou, Z.; Huang, Y.; and Wang, Y. 2022. " It's A Blessing and A Curse": Unpacking Creators' Practices with Non-Fungible Tokens (NFTs) and Their Communities. *arXiv preprint arXiv:2201.13233*.

Shen, H.; Knearem, T.; Ghosh, R.; Alkiek, K.; Krishna, K.; Liu, Y.; Ma, Z.; Petridis, S.; Peng, Y.-H.; Qiwei, L.; Rakshit, S.; Si, C.; Xie, Y.; Bigham, J. P.; Bentley, F.; Chai, J.; Lipton, Z.; Mei, Q.; Mihalcea, R.; Terry, M.; Yang, D.; Morris, M. R.; Resnick, P.; and Jurgens, D. 2024. Towards Bidirectional Human-AI Alignment: A Systematic Review for Clarifications, Framework, and Future Directions. 2406.09264.

Shneiderman, B. 2020a. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(4): 1–31.

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

The White House. 2023. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.

Tsai, L. L.; Pentland, A.; Braley, A.; Chen, N.; Enríquez, J. R.; and Reuel, A. 2024. Generative AI for Pro-Democracy Platforms. *MIT*.

Tseng, Y.-Y.; Sharma, T.; Zhang, L.; Stangl, A.; Findlater, L.; Wang, Y.; and Gurari, D. 2025. Biv-priv-seg: Locating private content in images taken by people with visual impairments. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 430–440. IEEE.

Tully, M.; Maksl, A.; Ashley, S.; Vraga, E. K.; and Craft, S. 2022. Defining and conceptualizing news literacy. *Journalism*, 23(8): 1589–1606.

Wallach, W.; and Marchant, G. E. 2018. An agile ethical/legal model for the international and national governance of AI and robotics. *Association for the Advancement of Artificial Intelligence*.

Wang, M.; Xing, J.; and Liu, Y. 2021. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*.

Wang, T.; Hayes, C. M.; and Bashir, M. 2022. Developing a Framework of Comprehensive Criteria for Privacy Protections. In *Future of Information and Communication Conference*, 905–918. Springer.

Weber, R. H. 2015. Realizing a new global cyberspace framework. *Normative Foundations and Guiding Principles*.

Weyl, E. G.; Ohlhaver, P.; and Buterin, V. 2022. Decentralized society: Finding web3's soul. *Available at SSRN 4105763*.

Willis, R.; Curato, N.; and Smith, G. 2022. Deliberative democracy and the climate crisis. *Wiley Interdisciplinary Reviews: Climate Change*, 13(2): e759.

Wojciech Zaremba. 2023. Democratic Input to AI. https://openai.com/index/democratic-inputs-to-ai/. Accessed: 2024-04-06.

Xu, B.; and Li, D. 2015. An empirical study of the motivations for content contribution and community participation in Wikipedia. *Information & management*, 52(3): 275–286.

Xu, C.; Zhang, J.; Chen, Z.; Xie, C.; Kang, M.; Potter, Y.; Wang, Z.; Yuan, Z.; Xiong, A.; Xiong, Z.; et al. 2025. MMDT: Decoding the Trustworthiness and Safety of Multimodal Foundation Models. *arXiv preprint arXiv:2503.14827*.

Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; and Schmid, C. 2021. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1686–1697.

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, A.; Walker, O.; Nguyen, K.; Dai, J.; Chen, A.; and Lee, M. K. 2023. Deliberating with AI: Improving Decision-Making for the Future through Participatory AI Design and Stakeholder Deliberation. 7: 1–32.

Zhang, B.; and Dafoe, A. 2020. US public opinion on the governance of artificial intelligence. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 187–193.

Zhang, B.; and Zhou, H.-S. 2017. Brief announcement: Statement voting and liquid democracy. In *Proceedings of the ACM Symposium on Principles of Distributed Computing*, 359–361.

Zheng, C.; Wu, Y.; Shi, C.; Ma, S.; Luo, J.; and Ma, X. 2023. Competent but Rigid: Identifying the Gap in Empowering AI to Participate Equally in Group Decision-Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–19.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.