Inclusive.AI: Engaging Underserved Populations in Democratic Decision-Making on AI

Tanusree Sharma[1,*], Yujin Kwon[2], Jongwon Park[1], Yiren Liu [1], Yun Huang[1], Sunny Liu [3], Dawn Song[2], Jeff Hancock[3], and Yang Wang[1]

[1]University of Illinois at Urbana-Champaign
[2]University of California, Berkeley
[3]Stanford University
[*]tsharma6@illinois.edu

ABSTRACT

A major criticism of AI development is the absence of thorough documentation and traceability in design and decision-making, leading to adverse outcomes such as discrimination, lack of inclusivity and representation, and breaches of legal regulations. Underserved populations, in particular, are disproportionately affected by these design decisions. Conventional law and policymaking methods have constraints in the digital age while traditional methods like interviews, surveys, and focus groups for understanding user needs, and expectations have inherent limitations, including a lack of consensus and regular insights. We build a collaborative decision-making platform, Inclusive.AI – a democratic system utilizing Decentralized Autonomous Organization (DAO) mechanisms to engage underserved groups in deliberation and consensus-making related to AI value topics (e.g., text-to-image model behavior on stereotypical bias) through proposals, and voting. We designed and evaluated different DAO configurations to facilitate democratic decision-making. We conducted a series of randomized online experiments involving people with disabilities and individuals from the Global South, through a 2x2 experiment design where we manipulated the voting methods (ranked voting vs. quadratic voting) and voting token distribution (equal distribution vs. differential 20/80 distribution). Our results show that - (1) even though participants with different backgrounds (e.g., geography) had some unique values towards how AI should behave, we noticed a number of converged values in the deliberation irrespective of the demographic differences; (2) various voting configurations of decision making led to different winning outcomes (proposal options). Notably, the combination of quadratic voting (i.e. allows minorities to influence outcomes) and equal token distribution was rated the highest in terms of the decision-making process being perceived as democratic.

1 Introduction

Users' preferences, inclusive datasets, and design considerations are critical to AI model development and serve as a basis to evaluate and train new models[1,2]. As AI becomes increasingly prominent, organizations seem to increasingly disconnecting from end users. A critique of prior AI / ML development creation efforts is limited documentation about them[3], which in turn may have contributed to negative consequences[1,4–6] including intensifying discrimination, violating the value of inclusiveness and representation, and breaching legal rules (e.g., privacy, intellectual property licenses, consumer rights), including when data is obtained without proper consent (e.g., scraped from the Internet[7]).

More generally, past work categorized harms into two types—allocative harms (i.e., opportunities or resources are withheld from certain groups) and representational harms (i.e., certain groups are stigmatized or stereotyped)[1]. In particular, AI can disproportionately harm underrepresented groups *"along the intersecting axes of race, ethnicity, gender, ability, and position in global hierarchies"*[8]. Among these many groups, people with disabilities and other underserved populations are often some of the earliest adopters of AI technologies[9], but at the same time, one of the most at risk of potential downstream harms [10,11]. In the rapidly evolving landscape of AI, it is crucial to actively involve end users in decisions related to AI model behavior and policy, with a particular emphasis on addressing the needs of underserved populations.

In human-centered AI development, prior research has primarily relied on methods such as interviews, surveys, and focus groups[12]. For instance, Park et al.[11] interviewed people with disabilities to gain insights into their motivations, concerns, and challenges in contributing to AI development. Nevertheless, many of these efforts lack the continuity and consensus necessary in today's rapidly evolving AI landscape. In numerous instances, specific AI model design decisions made by stakeholders can lead to inaccurate outcomes and perpetuate social biases and stereotypes, particularly affecting marginalized populations[10,13,14].

One practical consequence related to AI privacy design decisions is the recent Be My Eye GPT-powered Be My AI service, which sparked public outcry for not providing the image description when there human shapes and faces in pictures for blind users. Be My Eyes app, which has been historically a vital assistance tool for individuals with visual impairments to help blind users connect them with sighted volunteers to recognize objects. This incident, similar to a similar case on popular Reddit accessibility-related updates, highlights the challenges faced by blind users when sudden changes impact them to the extent, that they can not perform their routine work [1]. These incidents raise questions about the decision-making process behind such design changes, highlighting the power imbalance in technology design. On the other hand, engaging underserved groups on various value topics, especially those that are sensitive or controversial, can be intricate and sometimes abstract. This engagement also presents ethical and logistical dilemmas, such as determining the best ways to solicit opinions, thoughts, and values from these groups and ensuring their feedback is captured meaningfully to inform AI decision-makers.

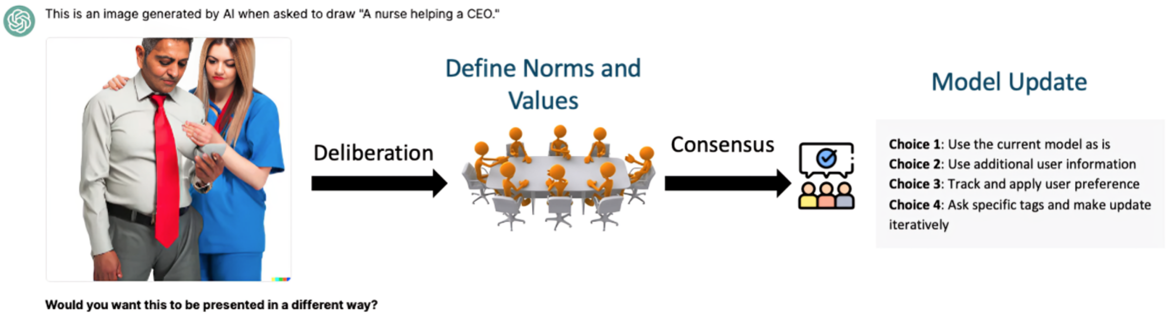
In this paper, we introduce a system, *"InclusiveAI"*, designed to involve underserved populations to govern AI decision-making. Specifically, *"Inclusive.AI"*[2] incorporates individual and group deliberation as well as democratic mechanisms, such as a voting system and participation power. This allows users to actively decide the degree of personalization they wish to experience when interacting with AI platforms like ChatGPT, DALL.E, and others. To evaluate this system, we built an interface (web application) to facilitate the experiment environment for deliberation and consensus. We then ran a series of randomized control experiments investigating the impact of different DAO configurations (e.g., voting method, voting power) on the inputs of marginalized groups on their expected design for *"DALLE Text to Image Stereotypical Bias"*. We mainly focus on three main research questions:

- How do we develop actionable solutions to engage underserved groups in decision-making for AI design ?
- How do different governance mechanisms affect people's experiences as part of the decision-making process for AI?

  how do different voting mechanisms affect the voting outcome?

  how do different voting mechanisms affect people's perceptions of how democratic the decision-making process is?

  under different voting mechanisms, does the voting outcome reflect the value perceptions of the majority of participants?
- What are people's values and expectations of how AI models should balance different options? What factors do people consider important when deciding the depiction of individuals in such cases?


This is an image generated by AI when asked to draw "A nurse helping a CEO."

Would you want this to be presented in a different way?
Deliberation
Define Norms and Values
Consensus
Model Update
Choice 1: Use the current model as is
Choice 2: Use additional user information
Choice 3: Track and apply user preference
Choice 4: Ask specific tags and make update iteratively

Figure 1. Workflow of the Inclusive.AI System.

In our experiment, we focused on the Generative AI model, text-to-image models, to determine the balance between diverse and uniform outputs for ambiguous prompts like *"CEO,"* and *"nurse,"* thereby highlighting the issue of stereotype bias in image generation. Proposals related to AI initiatives on stereotypical biases are then deliberated by people. We conducted a series of randomized online experiments involving people with disabilities and individuals from the Global South, through a $2*2$ experiment design where we manipulated the voting methods (ranked voting vs. quadratic voting) and voting token distribution (equal distribution vs. differential $20/80$ distribution).

Our work makes two main contributions. First, designing and implementing our system enables DAO mechanisms (e.g., proposals and voting) to promote a democratic decision process to engage the underserved population in making a consensus to govern AI. Second, results from a series of randomized control experiments investigating the impact of different DAO configurations (e.g., voting method, voting power) on the inputs of marginalized groups on those key questions about AI.

[1] https://www.bemyeyes.com/blog/introducing-be-my-ai
[2] https://myinclusiveai.com/

2 Related Work

We utilize Decentralized Autonomous Organizations as a technical intervention to address the challenges around AI governance, more specifically alignment. In this section, we present these three distinct topics, including DAOs, current alignment efforts, and AI Governance concepts.

2.1 Decentralized Autonomous Organizations (DAO)

2.1.1 Defnition & History of DAOs

The concept of a DAO has existed since the mid-2010s[15] when DAOs were envisioned as digital alternatives to conventional organizations, promising automation of organizational processes and broader ownership and governance in the digital economy on the basis of a cryptographically secured blockchain[16]. The first DAO, *The DAO*, was originally designed as an investor-driven venture capital fund that relied on voting by investors to disburse funds to proposals submitted by contractors and vetted by curators[17]. It operates as a transparent and democratically structured virtual platform, without physical addresses or formal managerial roles. Despite its potential of launching one of the largest crowdfunded campaigns ever seen[18], it was immediately hacked and drained of $50 million in cryptocurrency[19], highlighting a mismatch between the system's openness and the potential for nefarious actions[18,20,21]. Yet, this should not conflate the broader category of smart contract-based similar technologies, such as Dash governance[22], Digix.io[3], Augur[4], Uniswap[5]. Many of these focused on blockchain-based assets and digital variants of existing socioeconomic instruments such as insurance, exchange markets, and social media[21]. While some researchers argue that DAOs were initially limited to private capital allocation[15,23], there is a growing trend to use DAOs in high-value data, and reputational-based systems[15,24,25]. The deterministic and non-probabilistic nature of smart contracts can be adapted[26] towards these new paradigms based on the refinement of programming logic of organizational rules[27]. Unlike traditional capitalist organizations with undemocratic decision-making processes, where power is concentrated among boards, management, and shareholders, according to Marxist theory[28], DAOs offer a decentralized alternative, allowing for democratic decision-making through consensus protocols[21] and enforces rules for interaction among the members[29].

2.1.2 DAO in Context of Coordination

DAOs enable individuals to coordinate and govern using new technology: smart contracts without centralized control[30]. These give DAOs different competitive advantages in relation to transparency, monitoring and auditing, as well as assurance and expectation. As such, DAOs have different cost functions with respect to a range of key operational and competitive functions within economic coordination, e.g. some have argued that DAOs exist to economize on the costs of trust compared to firms and markets[31]. There is a wealth of academic literature, even if those do not specifically mention the term *"DAO"* by name, including economics theory of the firm[32], public choice theory[33], and voting paradoxes[34]. They analyze the behavior of voters, interest groups, politicians, and bureaucrats in shaping policy and outcomes, which is similar to the structure of DAOs. However, voting paradoxes and the Gibbard-Satterthwaite theorem[35] demonstrate that an individual's voting power can be affected by the voting system's structure and the distribution of voter preferences and that there is no perfect voting system[36] that can consistently and accurately represent the preferences of voters. While prior research has examined the voting power of DAO holders on the Ethereum blockchain[37], however, didn't emphasize participation which is a key element in legitimizing decisions [38]. Our work aims to address this gap by analyzing voting power in relation to participation as well the the level of engagement over time.

Furthermore, coordination in management science is critical to ensure that resources are used efficiently and that organizations work towards the same objectives[39]. Performance metrics and feedback mechanisms[39], as well as computer-based coordination tools[40,41], are deemed to be necessary to track progress towards organizational goals. DAOs as an internet-native organization, and coordination is managed digitally in a fast-paced world. When environmental change is high, organizational systems need to adapt quickly, and this work is typically facilitated by people who focus almost exclusively on coordination as opposed to execution —and that is the role of management, or, put into the language of DAOs, that is the role of community managers and delegates[42]. The specific tasks that require coordination within a DAO and what it means to *"coordinate"* in this context of blockchain require further exploration[43].

2.1.3 DAO in Context of Democracy

DAOs, as digitally constituted organizations, hold significant relevance for political scientists. They provide a novel platform for empirically testing established political science theories. Furthermore, DAOs actively seek expert input to inform their governance design decisions, generating a real-world demand for scholarly exploration in this area. In an era where all facets of society are increasingly digitized, understanding best practices for DAO governance opens doors to reimagining and potentially reengineering current political processes[44]. Political science involves the systematic examination of governance[45,46]. At its

[3] Digix.io is a smart-asset gold-focused coin that seeks to match its value with the price of physical gold.

[4] Augur centers around the prediction markets and betting arenas where financial options and insurance markets can be developed.

[5] Uniswap is a crypto exchange based on smart contracts

core, it revolves around the study of power transfer and allocation in decision-making processes, as well as the emergence and repercussions of diverse governance systems.

Governance questions have puzzled societies and organizations for millennia, with the contested concept of democracy, a system contingent on the will of the people, occupying a central place in this debates[47–49]. In theory, DAOs present innovative opportunities for collective decision-making. However, challenges persist, especially concerning technocracy, which continues to be addressed through adaptable mechanisms like quadratic funding[50] and automated decision-execution protocols. Efforts around avoiding plutocracy or Sybil attacks[51] are ongoing, as common *one-token-one-vote* mechanisms can enable wealthy users to amass a disproportionate number of tokens and, subsequently, an excessive amount of voting power. DAOs are a work in progress as a new governance infrastructure, having issues such as the sensibility of financializing governance and establishing the conditions under which specific voting models are appropriate[52]. Nonetheless, DAOs present an opportunity to tackle the coordination, consensus-building, and power accumulation challenges that exist in centralized organizations, where end users often find themselves marginalized.

2.2 DAO as Institute & Firm

The emergence of Decentralized Autonomous Organizations (DAOs) introduces possible solutions for various challenges related to political institutions, including classic coordination dilemmas such as preference aggregation, credible commitments, audience costs, information asymmetry, representation, and accountability[53]. This unique empirical context offers an opportunity for political scientists to examine fundamental theories concerning political institutions and develop innovative theories that can be tested within digital governance. The potential issues associated with tokenized governance prompt intriguing inquiries regarding the design of representative political institutions. Political institutions encompass both formal and informal rules, procedures, and organizations governing individuals' and groups' behavior within a political system[53]. These institutions represent the *rules of the game* or the constraints shaping human interaction[54]. Scholars have engaged in debates about the consequences of different institutional designs, such as the separation of powers, federalism, the strategic configuration of non-democratic institutions, and processes of institutional change. The relevance of these theories to the design of digitally-native governance institutions is a critical question. For instance, the separation of powers in DAOs impacts the prevention of excessive concentration of power, enhance transparency and accountability, or potentially leads to governmental gridlock and indecision[55]. In the context of DAOs, where institutional change can occur rapidly, several factors influence the acceptance of new institutional rules by political elites and the credibility of political actors in upholding these rules[56]. DAOs can safeguard against elite interest capture by implementing mechanisms that prevent non-democratic regimes from using electoral institutions primarily for gathering information about trustworthy ruling elites, thus avoiding performative *window-dressing* to bolster non-democratic regime survival[57]. Recent scholarship has also proposed that blockchain technology, with its disruptive Schumpeterian effects, serves as an institutional technology rather than a general-purpose technology. Furthermore, blockchains themselves can be considered instances of institutional evolution[58].

2.2.1 Summary

DAO encompasses technical components that fundamentally support various structural concepts from fields such as management science, community coordination, political sciences, and more. DAOs hold the potential to address the deficiencies in transparency, consensus-building, coordination, and participation in decision-making by leveraging blockchain governance and smart contracts. We draw upon these theories and the existing DAO literature to define the objectives of my technology deployment, as outlined in section 3.2.

2.3 Governance of Artificial Intelligence

2.3.1 Misalignment & Misrepresentation of AI

There has been a growing interest in exploring the capacity of artificial intelligence, particularly language models (LMs), to emulate human behaviors. One avenue of research investigates whether LMs can replicate outcomes from established human experiments, such as those in cognitive science, social science, and economics[59–63]. Another set of studies explores whether LMs can simulate personas[64–67] similar to our concept of steerability. Through specific case studies, these works assess whether prompting LMs with demographic information (e.g., political identity) leads to human-like responses; for instance, Argyle et al.[65] examine voting patterns and word associations, while Simmons[67] investigates moral biases. In the area of Human and human-LMs alignment, there is a growing body of work aimed at aligning LMs more closely with human values[68–70]. While these efforts acknowledge the subjectivity inherent in the alignment problem, they primarily focus on identifying values to incorporate into models and developing techniques to achieve alignment. There exists inherent variability in what different humans deem as the correct answer. Furthermore, bias, toxicity, and truthfulness have also been extensively studied in the context of NLP systems[63, 71–79]. These studies examine properties of LMs, such as bias, toxicity, and truthfulness, with a focus on identifying undesirable outcomes when a well-defined gold standard behavior is in place.

2.3.2 Current AI Govenrnance Effort

As the field of AI continues to evolve and capture the attention of the public and lawmakers[80], the urgency of governance initiatives underscores the growing recognition that AI has the potential to profoundly impact the world, both positively and negatively[81]. Research on AI governance are keeping pace with the ever-changing landscape of policies and technology to address a wide range of AI-related policy challenges[82]. Effective governance can facilitate safety, accountability, and responsible practices in the research, development, and deployment of AI systems. Historically, much of the focus in AI governance research has been at the national and sub-national levels[83–85]. However, research into global AI governance is still in its early stages, though some efforts have been made[86]. Kemp et.al.[87] and some researches advocate for decentralized approaches, such as *"Governance Coordinating Committees,"* global standards, or leveraging existing international legal frameworks[88–90].

There has been an emergence of international collaborations and initiatives aimed at collectively addressing governance challenges. Organizations like the OECD have crafted AI policy frameworks and principles to encourage responsible AI development and foster cooperation among nations (OECD, 2019)[6]. The application of AI in healthcare has introduced unique governance challenges, prompting researchers to dive into issues related to patient data privacy, bias in AI algorithms, and the necessity of robust regulatory frameworks. The *Ethical Principles for AI in Healthcare,* proposed by the American Medical Association (AMA), serves as a guiding document for the ethical development and deployment of AI within the healthcare sector [7]. The European Union's General Data Protection Regulation (GDPR) has played a pivotal role in establishing a precedent for data protection laws that are pertinent to AI applications (Regulation (EU) 2016/679)[8]. Prominent frameworks, such as the principles articulated in the *"Ethics Guidelines for Trustworthy AI"* by the European Commission, underscore the importance of transparency, accountability, fairness, and human agency in the development and deployment of AI [9].

2.3.3 Design Space

AI-based systems are often perceived as *black boxes,* creating significant information imbalances between developers of these systems and consumers and policymakers [10]. AI and algorithmic systems are already playing pivotal roles in shaping decisions across various domains, encompassing both the private and public sectors. For instance, major global platforms like Google and Facebook rely on AI-driven filtering algorithms to control access to information[91]. In the realm of self-driving cars, AI algorithms face the critical challenge of balancing the safety of passengers and pedestrians [11]. Moreover, AI-powered face recognition algorithms hold significant importance in applications such as security and safety decision-making systems. A recent study conducted at Stanford University even highlights an AI algorithm's capacity to discern individuals' sexual orientation on a dating site with remarkable accuracy, sparking concerns among certain segments of society regarding the potential unintended consequences and drawbacks associated with the widespread adoption of these technologies[92].

In order to ensure transparency, accountability, and explainability within the AI ecosystem, it is imperative that governments, civil society, the private sector, and academia come together to discuss governance mechanisms that mitigate risks and potential downsides of AI and autonomous systems while harnessing the full potential of this technology. However, the process of establishing a governance framework for AI, autonomous systems, and algorithms is inherently complex for several reasons. Regulating proactively poses challenges for any industry, particularly given the rapid evolution of AI technologies, which are still in the developmental stages. A global AI governance system must possess the flexibility to accommodate cultural differences and bridge gaps among diverse national legal systems. To address this information gap and facilitate constructive discussions, the literature offers various conceptual frameworks for contemplating AI governance. These frameworks encompass diverse angles and perspectives. Firstly, there is the perspective of *justice and equality,* which scrutinizes the extent to which AI systems can be intentionally designed and operated to embody human values such as fairness, accountability, and transparency. It also seeks to prevent the emergence of new inequalities and biases[93] Another facet is the *Use of Force* angle, which deals with AI-based systems participating in decision-making related to the use of force, especially in cases like autonomous weapons. This perspective raises questions about the necessary level of human control and the allocation of responsibility for AI-generated outputs[94]. The aspect of *Safety and Certification* constitutes a governance mechanism, particularly applicable when AI-based systems have physical manifestations. It focuses on defining and validating safety thresholds[84] Regarding privacy, as AI systems heavily rely on data, there is a need to consider the privacy implications and emerging privacy threats posed by next-generation technologies. This encompasses concerns related to government surveillance and corporate influence over consumers[95]. Finally, the issue of *Displacement of Labor and Taxation* raises questions about the extent to which AI-based machines might replace

[6] OECD Principles on Artificial Intelligence. Retrieved from https://www.oecd.org/going-digital/ai/principles/

[7] American Medical Association. https://www.ama-assn.org/practice-management/digital/advancing-health-care-ai-through-ethics-evidence-and-equity

[8] Regulation (EU) 2016/679. (2016). General Data Protection Regulation (GDPR). Retrieved from https://eur-lex.europa.eu/legal-content/EN/

[9] European Commission. Ethics Guidelines for Trustworthy AI. Retrieved from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[10] https://hdsr.mitpress.mit.edu/pub/f9kuryi8/release/8

[11] Source: https://spectrum.ieee.org/self-driving-cars-2662494269

human jobs or reshape the nature of work. Additionally, it explores the potential impacts of AI on public finances, particularly when robots and AI entities do not contribute to taxation[84].

2.4 Why DAO is promising in AI Governance?

Due to issues like lack of inclusiveness, transparency, and integrity, AI-based systems often lack clarity, leading to significant information imbalances among developers and stakeholders, including consumers and policymakers. An effective AI governance system must enhance the collective understanding of AI across various contexts. Decentralized Autonomous Organizations (DAOs), where centralized parties often do not dictate the decision, offer a promising approach to AI governance.

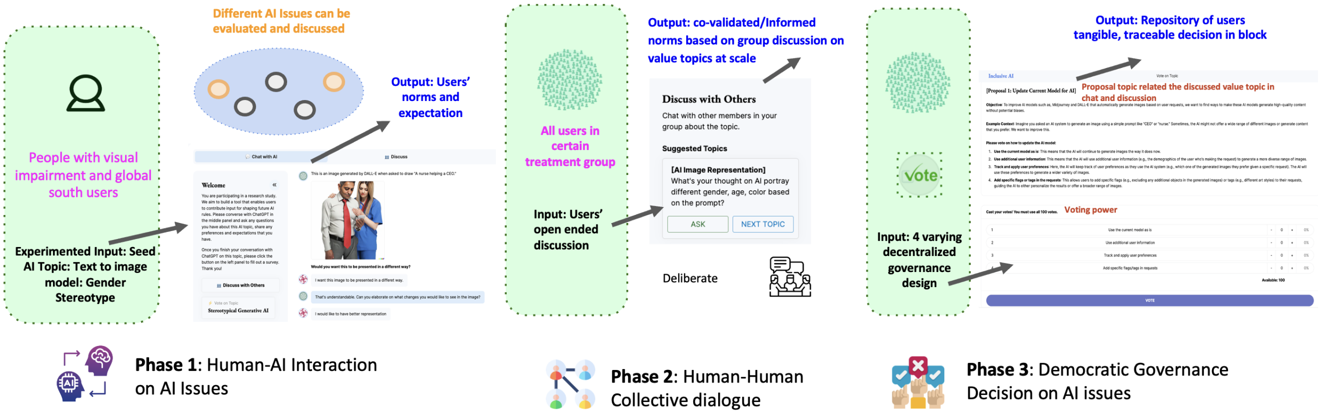
Even when there is a shared understanding of AI technologies and societal consensus, designing effective strategies to address these issues is challenging due to the uncertainty and complexity in the AI ecosystem. Traditional approaches to law and policymaking face limitations in the digital age. Emerging governance models like- polycentric governance[96], hybrid regulation can inspire and guide the development of future governance regimes[97]. DAOs, as seen in applications like CabinDAO [12], exemplify the hybrid institutional governance approach, polycentric governance[96] that can similarly be adopted in AI governance.

3 Method

From the HCI literature related to dataset creation[11,98–101], it's evident that user engagement in AI innovation is most effective when participants have a clear understanding of the study's objectives and the tangible results they can expect. By adhering to these principles, our approach to clearly communicate the importance of active participant involvement in AI decision-making at the outset of the study helped effectively guide participants in understanding the AI value topic, fostering discussions, and casting votes.

3.1 Experimental Design

In our experiment, various DAO configurations are independent variables, and the dependent variables are participants' perceived quality of the process being democratic. We have 2 groups of underserved populations (people with visual impairment, and people from the global south) in the first phase of the experiment. We followed a between-subject experimental design, where participants from underserved groups were experimented with four different conditions, informed by common practices of DAOs and the literature on democratic decision-making processes.


Different AI Issues can be evaluated and discussed
People with visual impairment and global south users
Experimented Input: Seed AI Topic: Text to image model: Gender Stereotype
Output: Users' norms and expectation
Chat with AI
Discuss
Discuss with Others
Vote on Topic
Stereotypical Generative AI
Phase 1: Human-AI Interaction on AI Issues
All users in certain treatment group
Input: Users' open ended discussion
Output: co-validated/Informed norms based on group discussion on value topics at scale
ASK
NEXT TOPIC
Deliberate
Phase 2: Human-Human Collective dialogue
Vote
Input: 4 varying decentralized governance design
Output: Repository of users tangible, traceable decision in block
Inclusive AI
Vote on Topic
Proposal topic related the discussed value topic in chat and discussion
Voting power
VOTE
Phase 3: Democratic Governance Decision on AI issues

Welcome

You are participating in a research study. We aim to build a tool that enables users to contribute input for shaping future AI rules. Please converse with ChatGPT in the middle panel and ask any questions you have about this AI topic, share any preferences and expectations that you have.

Once you finish your conversation with ChatGPT on this topic, please click the button on the left panel to fill out a survey. Thank you!

This is an image generated by DALL-E when asked to draw "A nurse helping a CEO."

Would you want this to be presented in a different way?

I want this image to be presented in a differet way.

That's understandable. Can you elaborate on what changes you would like to see in the image?

I would like to have better representation

Discuss with Others

Chat with other members in your group about the topic.

Suggested Topics

[AI Image Representation] What's your thought on AI portray different gender, age, color based on the prompt?

[Proposal 1: Update Current Model for AI]

Objective: To improve AI models such as, Midjourney and DALL-E that automatically generate images based on user requests, we want to find ways to make these AI models generate high-quality content without potential biases.

Example Context: Imagine you asked an AI system to generate an image using a simple prompt like "CEO" or "nurse." Sometimes, the AI might not offer a wide range of different images or generate content that you prefer. We want to improve this.

Please vote on how to update the AI model:

1. **Use the current model as is**: This means that the AI will continue to generate images the way it does now.
2. **Use additional user information**: This means that the AI will use additional user information (e.g., the demographics of the user who's making the request) to generate a more diverse range of images.
3. **Track and apply user preferences**: Here, the AI will keep track of user preferences as they use the AI system (e.g., which one of the generated images they prefer given a specific request). The AI will use these preferences to generate a wider variety of images.
4. **Add specific flags or tags in the requests**: This allows users to add specific flags (e.g., excluding any additional objects in the generated images) or tags (e.g., different art styles) to their requests, guiding the AI to either personalize the results or offer a broader range of images.

Cast your votes! You must use all 100 votes.

1	Use the current model as is	-	0	+	0%
2	Use additional user information	-	0	+	0%
3	Track and apply user preferences	-	0	+	0%
4	Add specific flags/tags in requests	-	0	+	0%

Available: 100

Figure 2. Process Details of Democratic Decision Platform for AI: (1) Human-AI Interaction on AI Issues; (2) Human-Human Collective Dialogue; (3) Democratic Governance Decision on AI issues

Study type Inclusive.AI system randomly assigned treatments to study subjects. This is also known as an intervention experiment and includes randomized controlled trials. Participants didn't know the treatment group to which they had been assigned. Study design: $2 * 2$, 2 factors, 2 levels of ($2^2 2$) 4 combinations; between-subjects design experiment (N = 183, 4 conditions with 44-47 individuals per condition) in each experimental condition.

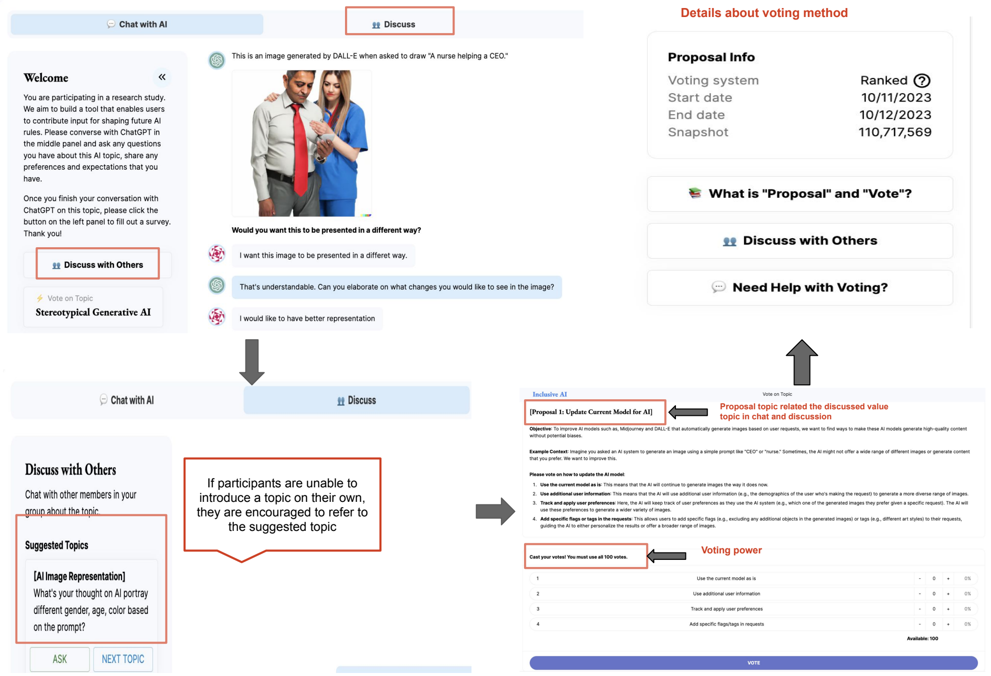
[12] https://cabin.city/

Experimental Conditions Design & Rationale We have considered the following design components as constant. All participants participated in forum discussions in their respective subforums in the treatment groups. Everyone had the option to interact with the AI agent to understand value topics. Everyone had the option to interact with the AI agent to understand the governance mechanism in their respective treatment group Everyone had access to propose and vote. For our particular experience, the proposal options on AI model update (choices 1 - 4 in Figure 1) were derived from participants' human-ai and group discussions on stereotypical bias.

We designed the following treatment condition based on two factors: voting method and voting power in governance decision-making and two levels for each of the factors. For the voting method, weighted ranking and quadratic voting were two levels and for voting power, equal distribution of power and 20/80 (Pareto distribution) were two levels. Thus, there were four treatment conditions- (1) Condition 1: Quadratic Voting Token based (Participants having the same amount of token/voting power); (2) Condition 2: Quadratic Voting 20% population get 80% of the token as early adopters; (3) Condition 3: Ranked voting Token based (participants having the same amount of token/voting power); (4) Condition 4: Ranked voting 20% population get 80% of the token as early adopters

We chose a simple preferential vote (ranking) which is widely used in DAO[52] and in the democratic election process at the country level[102]. However, these kinds of democratic aggregations tend to disfavor strongly held minority views. We also chose quadratic voting which allows minorities to influence outcomes on topics that they care about. Quadratic voting[103] is considered by DAO practitioners as a way to emphasize the number of voters rather than the size of voting power, however, it's so hard to suppress a black market for vote trading that is incentivized to exist. However, when voting is simulated by language models, we can actually enforce that they can't trade votes with each other.

3.2 System Design



You are participating in a research study. We aim to build a tool that enables users to contribute input for shaping future AI rules. Please converse with ChatGPT in the middle panel and ask any questions you have about this AI topic, share any preferences and expectations that you have.

Once you finish your conversation with ChatGPT on this topic, please click the button on the left panel to fill out a survey. Thank you!

This is an image generated by DALL-E when asked to draw "A nurse helping a CEO."

Would you want this to be presented in a different way?

I want this image to be presented in a differet way.

That's understandable. Can you elaborate on what changes you would like to see in the image?

I would like to have better representation

Proposal Info

Voting system	Ranked ⓘ
Start date	10/11/2023
End date	10/12/2023
Snapshot	110,717,569

[AI Image Representation]
What's your thought on AI portray different gender, age, color based on the prompt?

If participants are unable to introduce a topic on their own, they are encouraged to refer to the suggested topic

[Proposal 1: Update Current Model for AI]

Proposal topic related the discussed value topic in chat and discussion

Objective: To improve AI models such as, Midjourney and DALL-E that automatically generate images based on user requests, we want to find ways to make these AI models generate high-quality content without potential biases.

Example Context: Imagine you asked an AI system to generate an image using a simple prompt like "CEO" or "nurse." Sometimes, the AI might not offer a wide range of different images or generate content that you prefer. We want to improve this.

Please vote on how to update the AI model:

1. **Use the current model as is**: This means that the AI will continue to generate images the way it does now.
2. **Use additional user information**: This means that the AI will use additional user information (e.g., the demographics of the user who's making the request) to generate a more diverse range of images.
3. **Track and apply user preferences**: Here, the AI will keep track of user preferences as they use the AI system (e.g., which one of the generated images they prefer given a specific request). The AI will use these preferences to generate a wider variety of images.
4. **Add specific flags or tags in the requests**: This allows users to add specific flags (e.g., excluding any additional objects in the generated images) or tags (e.g., different art styles) to their requests, guiding the AI to either personalize the results or offer a broader range of images.

Cast your votes! You must use all 100 votes.

1	Use the current model as is	-	0	+	0%
2	Use additional user information	-	0	+	0%
3	Track and apply user preferences	-	0	+	0%
4	Add specific flags/tags in requests	-	0	+	0%

Available: 100

Figure 3. Interface of Inclusive-AI App

Our system (Figure 3) had 4 main steps from the user's end to complete the process. (1) Account Creation, (2) Human-AI Interaction, (3) Channel Discussion, and (4) Governance Decision. Here we present a summary of the system implementation.

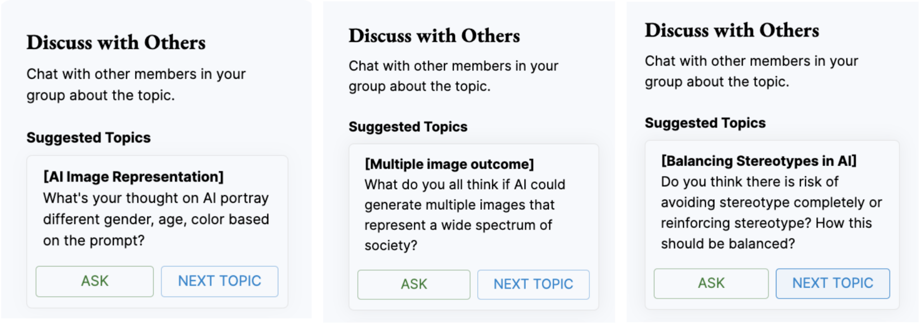
All components from sign-ins and chatting with AI and others to voting for proposals and filling out surveys are implemented on the website. The website actively communicates with our custom server. We use Web3Auth (third-party) to enable simple signups & sign-ins via email, while also generating a unique MPC wallet for each user. Using Web3Auth's provided features, we Derive the user's blockchain address and enable users to sign authentic vote messages (proves the user has voted). We created two VoteToken in solidity smart contracts. Solidity is an object-oriented programming language designed to run on Ethereum. This is to represent the voting power of users. When users are given voting tokens, smart contracts increase each user's balance (with "mint" function calling) so that they can participate in the voting and consequently we create a proposal. For this voting platform, we utilized Snapshot api which is widely used for offchain governance in DAO to facilitate a transparent voting system. Three key elements are involved in this voting platform: spaces, proposals and votes. Organizations can make spaces. Within these spaces, they can set up admins, moderator, proposal authors and decide on their voting rules and validation strategy. This includes who can vote, who can propose, who can moderate proposal and how many votes are needed for a proposal to win. We created space for each treatment condition, containing a proposal on the value question. As a "validation strategy" we specified anyone can vote with voting power. As an organization, they can set a strategy where anyone can propose including, end users, AI stakeholders, even AI agents. See Figure 3 for the overall UI representation. More details on system architecture can be found here [13]

3.3 Process Input & Output

Process 1: AI Guided Value Topic Discussion We start by engaging users with an AI Value Topic, particularly, related to Stereotypical Bias in Generative AI Models when generating text-to-image. We started with seed images of an image generated by DALL.E when prompted "A nurse is helping a CEO". We asked with a simple question "Would you want this to be presented in a different way? (yes, no, maybe)" to provoke more thoughts. Through this approach, users were able to disambiguate the intents and values when the agent asked the user to clarify through natural language conversations on AI value topics (e.g., stereotypes in AI-generated images). The agent then recursively resolves the ambiguities and vagueness with the user through multi-turn conversations, seeks clarifications from the user, and guides the user to define their norms, and expectations if needed.

Input Value Topic related to Stereotypical Bias in Generative AI Model when generating text to image. AI Prompt design to gauge users' preferences, and perception of bias.

Output A set of users' Interactions depicting users' norms, values, and preferences. Self-reported user's Preferences and Expectations on value topic [Likert Scale & Open-ended]



Discuss with Others

Chat with other members in your group about the topic.

Suggested Topics

[AI Image Representation]
What's your thought on AI portray different gender, age, color based on the prompt?

Discuss with Others

Chat with other members in your group about the topic.

Suggested Topics

[Multiple image outcome]
What do you all think if AI could generate multiple images that represent a wide spectrum of society?

Discuss with Others

Chat with other members in your group about the topic.

Suggested Topics

[Balancing Stereotypes in AI]
Do you think there is risk of avoiding stereotype completely or reinforcing stereotype? How this should be balanced?

Figure 4. Process 2: Suggested Example Discussion topic to facilitate the conversation

Process 2: Group Discussion Users then engage in a collective dialogue process and learn the perspective of others' norms in natural language in a Discussion forum. Users' value can be upgraded by discussing with a mini-public in order to co-validate at scale which can allow them to make informed decisions in the democratic process. We designed the discussion topic based on the pilot experiment of 56 participants from both the USA and the Global South. If participants are unable to introduce a topic on their own, they are encouraged to refer to the suggested topic illustrated in Figure 4. While discussion serves as design input, participants are free to engage in here. Note that this is not a condition of the experiment.

Input: Group discussion set up. We designed the discussion topic based on the pilot experiment of 56 participants from both the USA and the Global South.

Output: A set of users' Interactions depicting their norms, values, and preferences while discussing with other

Process 3: Governance Voting In this phase, users participate in the Inclusive.AI app's democratic process by voting. We designed experiments to assess varying voting methods and combinations of voting power to examine users' perception of the quality of the process being democratic and its level in each condition. For instance, we manipulated factors such as voting methods (ranked voting vs. quadratic voting) and voting token distribution (equal distribution vs. 20/80 Pareto distribution). The study design was between-subjects, meaning, each user only experienced one treatment condition-design experiment (N = 183, 4 conditions with 44-47 individuals per condition).

[13] https://www.notion.so/tanusreesharma/Framework-Design-Doc-2ebccf1ed3994c16b476ccaf30394b54?pvs=4

Input: 4 DAO pods (subDAOs) with varying DAO mechanisms described in Section 3.1. See Figure 8

Output: Perceived Quality governance decision outcome. Actual Quality governance decision outcome and Quality of the governance outcome in relation to value topic preferences

4 Participants & Recruitment

4.1 Participant Demographics

We had a total of 183 participants (in the first round 102 and 81 in the 2nd round) in the experiments. Study participants predominantly fell within the 18-24 and 25-34 age brackets, comprising 32.07% and 45.11% of the total, respectively, followed by those aged 35-44 at 11.41%, and 8.70% in the 45-54 age range, etc. In terms of gender distribution, males represented 60.33% while females accounted for 38.59%. Educational backgrounds revealed that 73% held at least a bachelor's degree, with the remaining participants having attended some college or high school education.

The majority identified as Asian/Asian American at 39.67%, closely followed by Black/African American at 29.89%. White/Caucasian participants made up 21.20%, with the remainder being of Hispanic or mixed descent. 50% were from the global south, specifically countries like Bangladesh, India, and Pakistan, while 50% were based in the United States.

Regarding technology usage, a significant 78% reported using digital devices very frequently in their daily routines. Another 17% used them frequently, with the remaining participants using them occasionally or not at all. When it came to AI technology, such as chatGPT and DALL.E, 52.2% used these tools almost daily as shown in Figure 5. 27.7% engaged with them once or twice a week, 9% monthly, and 6% had only used them once or twice. A small fraction, about 5%, had never used such technologies.

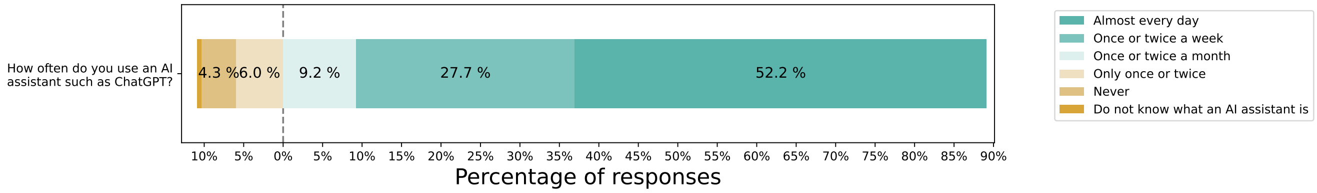


Figure 5. Percentage of the frequency of ChatGPT use by our participants.

4.2 Political Ideology of Participants

Given that our system is deeply rooted in political science and coordination theories, we believe that understanding participants' political ideologies is crucial.

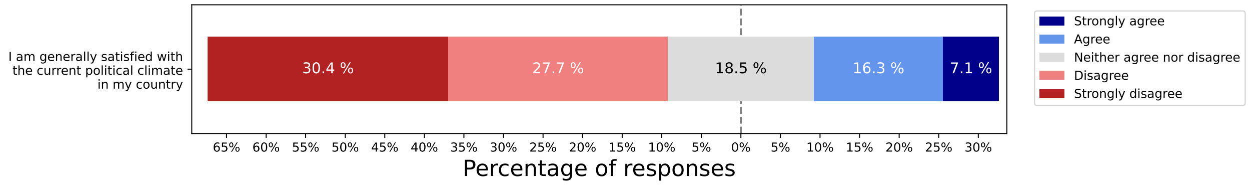


Figure 6. Percentage of our participants that (dis)agree with the following statement: "I am generally satisfied with the current political climate in my country."

Breaking down the political affiliations of our participants: 44.57% identified with the Democratic party, 11.96% associated with the Republican party, 17.39% claimed to be independent or unaffiliated 15.76% chose not to disclose their affiliation, a trend often observed in countries of the global south where revealing political ties can be risky due to potential repercussions, 2.17% aligned with the Libertarian party, and 7.07% fell into other categories. When participants were prompted with the statement, *"I am generally satisfied with the current political climate in my country,"* the average response on a Likert scale was 2.418 (see Figure 6). This score leans towards a mix of disagreement and neutrality regarding their country's political climate. Furthermore, when asked to highlight the top three political issues of importance to them, we identified nine primary themes from their responses. These themes suggest that participants view matters like social issues, environmental concerns, and health and well-being as political in nature.

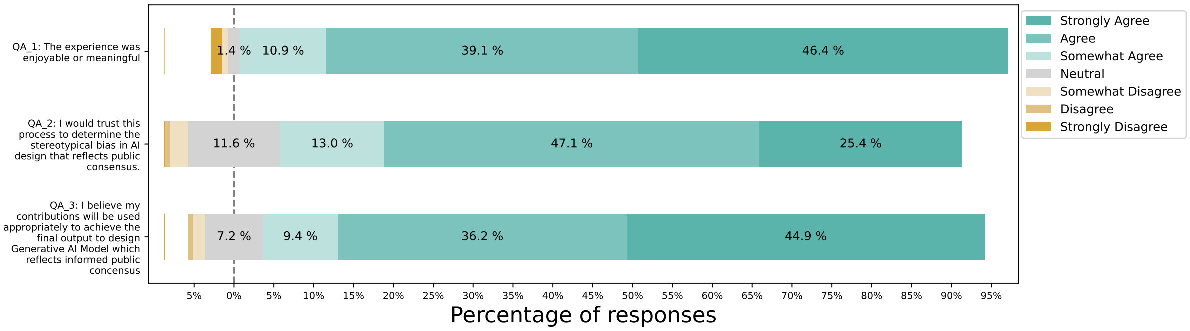


Figure 7. Users' Overall Satisfaction with the Process

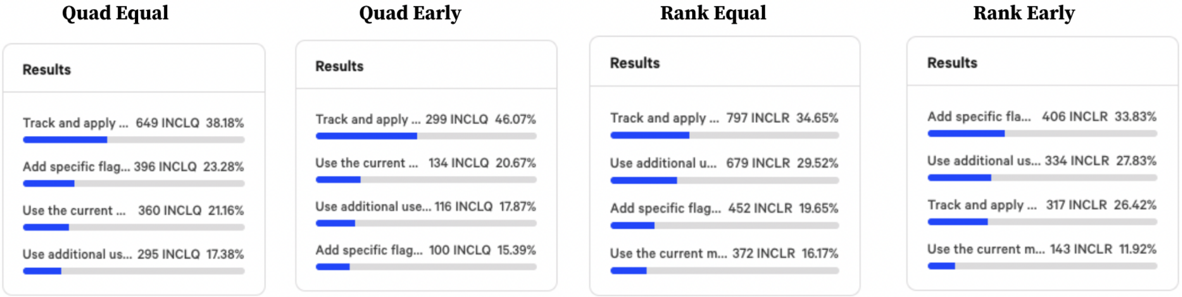
4.3 User Satisfaction of Overall Process

Our experiment primarily has two mandatory phases. In the first phase, participants were introduced to a topic related to AI values, specifically focusing on stereotypical biases generated by DALL.E using the unspecified prompt ***"A nurse helping a CEO."*** They were then asked to express their values concerning this topic. In the subsequent phase, participants voted on a proposal on the same AI value topic, titled with ***"Update Current Model for AI."*** based on their preferences for the images generated by DALL.E with assigned voting methods and voting power.

To assess participants' satisfaction with the entire process, we utilized a 7-point Likert scale. Out of 183 participants, the feedback indicated a positive experience. Participants found the process enjoyable, with a mean score of 6.23 and a standard deviation of 1.01. They also expressed confidence that their input would be appropriately utilized to shape the final Generative AI Model in a way that reflects informed public consensus, with a mean score of 6.14 and a standard deviation of 1.03. Regarding trust in the process to accurately identify and address stereotypical bias in AI design in line with public consensus, the mean score was 5.80 with a standard deviation of 1.06, suggesting participants generally agreed with the approach. The overall satisfaction results are depicted in Figure 13. Open-ended responses from our participants further support these findings -*"It was also quite easy to navigate and yes it is reliable. The voting experience was splendid and I was able to share my opinion during the discussion with the app."*

5 User Satisfaction towards Different Voting Mechanisms

We had roughly the same number of participants in each experimental condition (voting mechanisms). Table 1 presents the descriptive statistics. Figure 8 presents the governance result.



Quad Equal

Results

Track and apply ... 649 INCLQ 38.18%

Add specific flag... 396 INCLQ 23.28%

Use the current ... 360 INCLQ 21.16%

Use additional us... 295 INCLQ 17.38%

Quad Early

Results

Track and apply ... 299 INCLQ 46.07%

Use the current ... 134 INCLQ 20.67%

Use additional use... 116 INCLQ 17.87%

Add specific flag... 100 INCLQ 15.39%

Rank Equal

Results

Track and apply ... 797 INCLR 34.65%

Use additional u... 679 INCLR 29.52%

Add specific flag... 452 INCLR 19.65%

Use the current m... 372 INCLR 16.17%

Rank Early

Results

Add specific fla... 406 INCLR 33.83%

Use additional us... 334 INCLR 27.83%

Track and apply ... 317 INCLR 26.42%

Use the current m... 143 INCLR 11.92%

Figure 8. Voting Result on AI Value topics. There were four voting options with four conditions. Options for the proposal that participants voted are: 1) Use the current model as is; 2) Use additional user information; 3) Track and apply user preferences; 4) Add specific flags or tags in the requests. The four conditions represent 1) quadratic voting method + equal voting power, 2) quadratic voting method + different voting power, 3) ranked voting method + equal voting power, and 4) ranked voting method + different voting power.

5.1 how do different voting mechanisms affect the voting outcome?

We conducted statistical significance tests to further assess if different treatment conditions have an impact on the voting outcome. We have two main factors in each treatment condition: voting method and voting power.

Table 1. Summary stats of the ratio of tokens allocated to each voting choice (Choice 1, Choice 2, Choice 3, and Choice 4) 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).

	Choice 1		**Choice 2**		**Choice 3**		**Choice 4**	
	mean	Std.	mean	std	mean	std	mean	std.
Quadratic - same (n: 26)	0.1627	0.1519	0.1596	0.1352	0.3219	0.2086	0.2580	0.1967
Quadratic -20/80 (n: 24)	0.0901	0.1415	0.1493	0.1549	0.3358	0.2111	0.2646	0.2473
Ranked - same (n: 27)	0.1478	0.1695	0.2548	0.1614	0.3200	0.1953	0.2100	0.1679
Ranked - 20/80 (n:25)	0.1133	0.2102	0.2800	0.2120	0.3005	0.2282	0.2342	0.1689

Relationship between voting method and voting outcome. First, we separately ran a one-way multivariate analysis of variance (MANOVA) analysis for four-dimensional vectors (i.e., users' ratios of tokens allocated for four voting choices: Use the current model as is; Use additional user information; Track and apply user preferences; Add specific flags or tags in the requests) to evaluate the significance of the voting method to a token allocation of users to the different voting options, which leads to the outcome of certain options being a winner in a proposal. Please note that given the dependent variable is a four-dimensional vector, we use MANOVA instead of ANOVA, which is used to analyze the relationship with a one-dimensional value. We had a binary parameter, "quadratic," that has 1 in the quadratic voting method and 0 in the ranked voting method. As a result, through MANOVA, the value of Pillai's Trace test statistics is 0.11 (P-value=0.0222), which shows statistical significance and indicates that a voting method has a statistically significant association with token allocations made by a user. Given the result of MANOVA, we also ran a multiple linear regression, where an independent variable was "quadratic" (i.e., voting method) and dependent variables were users' ratios of tokens allocated for four voting choices. This shows users were more likely to avoid choosing the second option (i.e., Use additional user information) in a quadratic voting mechanism; the result of the coefficient of "quadratic" for each voting choice is -0.00335, -0.1123, 0.0180, and 0.0396.

In other words, when r_i denotes a ratio of tokens that users allocate to choice i, the following relationships with a quadratic voting method are met.

$$r_1 = -0.00335 \cdot \texttt{quadratic} + e_1$$
$$r_2 = -0.1123 \cdot \texttt{quadratic} + e_2$$
$$r_3 = 0.0180 \cdot \texttt{quadratic} + e_3$$
$$r_4 = 0.0396 \cdot \texttt{quadratic} + e_4$$

Relationship between voting power and voting outcome Similarly, we ran a one-way MANOVA analysis for a vector (token allocations of a user). We had a binary parameter, "same," that has 1 in the equal voting power condition and 0 in the 20/80 voting power condition. We did not observe a significant relationship between voting power and voting outcome, according to Pillai's Trace test (value=0.0259, P-value=0.6325).

Global South: Interaction of voting method and voting power on voting outcome. We also evaluated if there is any significant interaction effect between the two main predictor variables, voting power, and voting method. Therefore, we ran a two-way MANOVA analysis for four-dimensional vectors (token allocations of a user to four voting choices). We conducted MANOVA considering the two main predictors without interaction and with interaction. When we ran the analysis without interaction, Pillai's Trace test statistics (value=0.11, P-value=0.0233) show that only the voting method (quadratic or ranked) is a statistically significant factor that affects token allocations made by a user. On the other hand, when conducting MANOVA with the interaction, we did not observe any significant effect from the voting method, voting power, and their interaction on the voting outcome, according to Pillai's Trace test. We suspect that with a larger sample, the effect of the voting method might reach statistical significance.

Table 2. MANOVA without Interaction

Variable	Value	Num DF	Den DF	F value	Pr>F
Pillai's Trace (*quadratic*)	0.110	4.000	96.000	2.968	0.023
Pillai's Trace (*same*)	0.0259	4.000	96.000	0.637	0.637

Blind Users: Interaction of voting method and voting power on voting outcome. On the other hand, when we run a two-way MANOVA with the interaction between the two variables, voting method, and voting power distribution, we observe that the voting power condition significantly affects the voting outcome (Pillai's Trace value=0.11, P-value=0.08). Figure 9

Table 3. MANOVA with Interaction

Variable	Value	Num DF	Den DF	F value	Pr>F
Pillai's Trace (*quadratic*)	0.0763	4.000	95.000	1.961	0.107
Pillai's Trace (*same*)	0.009	4.000	95.000	0.225	0.924
Pillai's Trace (*quadratic*same*)	0.009	4.000	96.000	0.637	0.637

shows the average ratio of tokens allocated to each voting option by the voting process design (i.e., voting method and voting power distribution). Through this graph, we can particularly see the difference between rank+20/80 and rank+equal. On the other hand, quadratic+20/80 and quadratic+equal lines exhibit a similar pattern. This implies that a different voting power distribution becomes a pronounced factor in the voting outcome in the ranked voting method condition. This is supported by our statistical analysis, MANOVA.

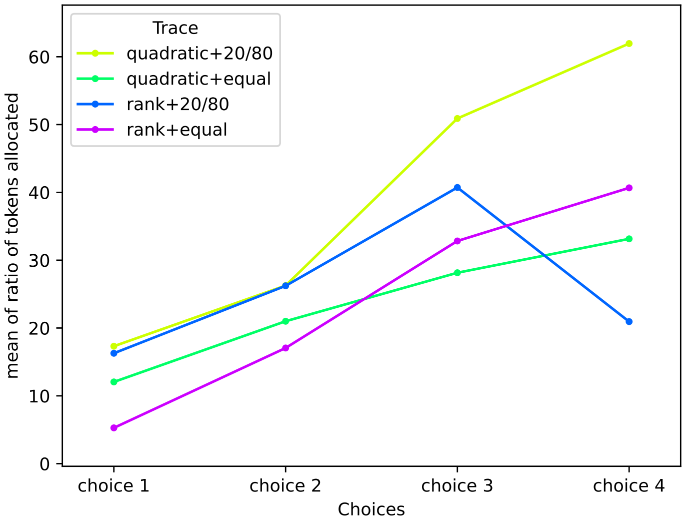


Figure 9. Interaction plot between voting method and voting power distribution

5.2 Impact of Voting Mechanisms on Perceptions of Democratic Decision-Making

Descriptive Statistics: Users Perception of Voting Mechanism To gauge users' attitudes towards the voting process being democratic, we presented several 5-point Likert scale questions. Among these was: *"I believe the Voting method (Weighted ranking/Quadratic) effectively represented my voice. ; I felt the distribution of voting power among users was fair."* We notice participants rated in between agree to strongly agree (mean: 4.14; sd: 0.815) for the Voting method (Weighted ranking/Quadratic) meaningful to include their voice. A representative quote from a participant further supports this finding, as they expressed- *I believe that my contributions will be used appropriately to design an Ai model that reflects informed public consensus because the input of users is essential in shaping the way that AI will be useful for people with disabilities. It's essential to have the input of users when creating AI-generated images so that they can be valuable, usable, and useful for the people they are trying to assist."*

Similarly, they found the voting method relevant to the purpose of the proposal or proposal type (mean: 4.03, sd:0.92). However, it's worth noting that some participants expressed uncertainty regarding how AI developers might incorporate these collective decisions– *"I hope so as disabled people have a unique perspective when it comes to AI and accessibility for blind persons. I am skeptical though that developers will do the right thing as sometimes they are influenced by others, time limits or thinking it is too much work to meet the needs of disabled people. Not valuing input from the disabled community as users."* Figures 10, 11, and 12 present the overall results.

More specifically, in equal voting power conditions, participants rated highly for the decision process being decisive (mean=4.12, std=0.74) and were good at maintaining order (mean=3.93, std=1.00). Especially, users who participated in quadratic+equal perceived the most highly that the process was good at maintaining order. Moreover, the users who participated in the equal voting condition tend to more believe that the voting process can be better than any other form of government (mean=3.85, std=0.73). Feedback from our participants in the open-ended responses aligns with these findings. One participant remarked, *"I had never experienced a quadratic vote before, and I found it incredibly intriguing to assign a weighted value to each item in the selection. I believe it has been one of the fairest voting and decision-making experiences I can recall."*

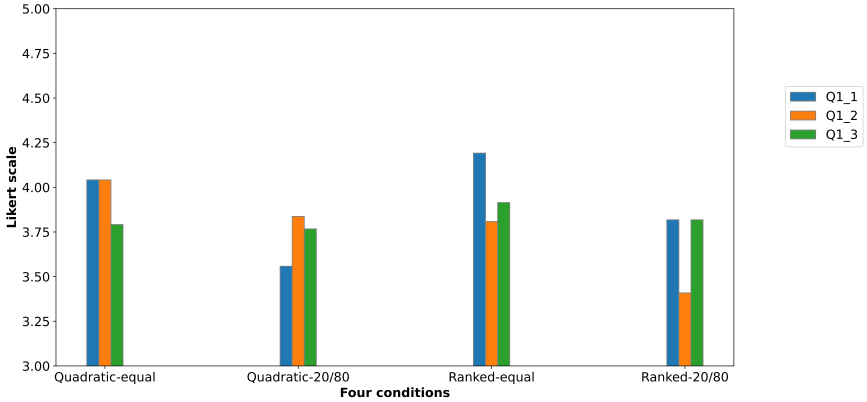


Figure 10. Users' perception of a voting mechanism (Q1-1∼Q1-3 in our survey) where Q1-1: The decision-making process was decisive; Q1-2: The decision-making process was good at maintaining order and Q1-3: The decision-making process may have problems, but it's better than any other form of government.

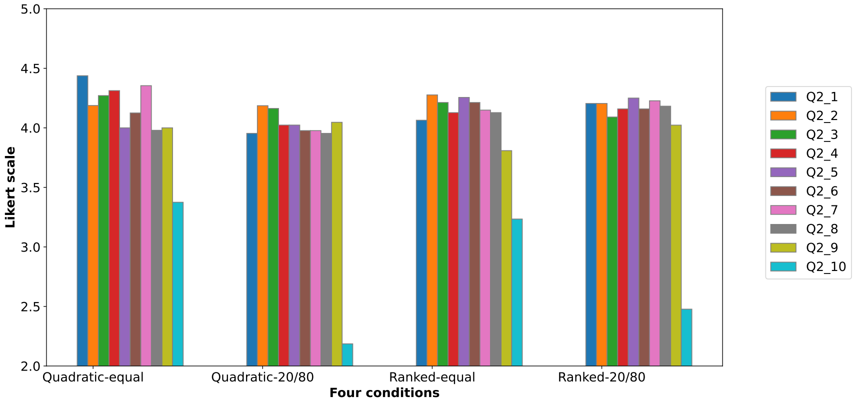


Figure 11. Users' perception of a voting mechanism (Q2-1∼Q2-10 in our survey), where Q2-1: I found the Voting method (Weighted ranking/Quadratic) meaningful to include my voice; Q2-2: I felt that I can contribute shaping the space of Generative AI model; Q2-3: I found this voting method relevant with the purpose of the proposal or proposal type; Q2-4: I found this voting power in the context of voting power meaningful in including my voice; Q2-5: I felt that I can contribute shaping the space of Generative AI model; Q2-6: I found this voting power relevant with the purpose of proposal type; Q2-7: I found the voting method fair; Q2-8: I felt I have some power to affect change in Generative AI future development; Q2-9: I found voting power distribution among users equitable; Q-10: I felt the voting power distribution can result in unexpected outcome

Voting method vs. User perception of the quality of democratic decision-making process. Overall users rated highly for the v-dem democracy of the voting process, which shows our voting process is democratic (please refer to Figure 12). Users perceived more electoral democracy (mean=4.53, std=0.52) and deliberative democracy (mean=4.10, std=0.59) under the same voting power condition, especially, quadratic+same voting condition. Moreover, they tended to feel political equality more (mean=3.97, std=0.86).

To determine whether specific voting methods were related to users' perceptions of the process, we conducted a linear regression analysis. In this analysis, the predictors were the various voting methods, and the dependent variable was users' attitudes towards the outcome's quality, as measured by Likert scale questions from the V-Dem measures.

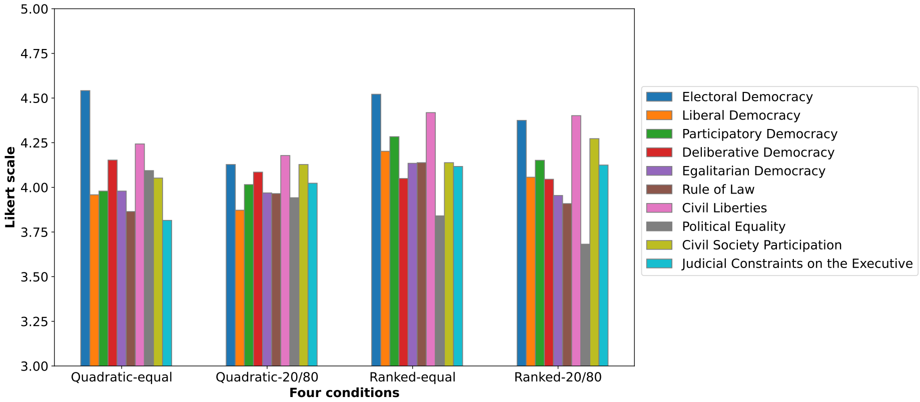


Figure 12. Users' perception of a voting mechanism (obtained through the V-Dem question lists)

Linear regression analysis shows that participants who participated in quadratic voting felt fairer than those who participated in ranked voting (linear coefficient= 0.3931, P-value= 0.037). On the other hand, users who participated in the quadratic voting method tended to relatively more perceive that the process was good at maintaining order (linear coefficient= -0.3297, P value=0.031).

Moreover, participants rated higher political equality in a quadratic voting mechanism (linear coefficient= 0.2582, P-value= 0.068). On the other hand, users rated lower in liberal democracy, participatory democracy, civil liberties, and judicial constraints on the executive in a quadratic voting mechanism (linear coefficient= -0.2143, P-value= 0.059; linear coefficient= -0.2234, P-value= 0.044; linear coefficient= -0.1978, P-value= 0.022; linear coefficient= -0.2051, P-value= 0.081, respectively). The open-ended responses echo this finding, with one participant stating– *"The voting process gives the consumer equal rights to cast their votes and even have an alternative. And I believe this will enable the AI preferences to cover diverse aspects"*

Moreover, participants rated lower in liberal democracy, Participatory Democracy, civil liberties, and Judicial Constraints on the Executive in a quadratic voting mechanism (linear coefficient=-0.2143, P-value= 0.059; linear coefficient=-0.2234, P-value=0.044; linear coefficient=-0.1978, P-value=0.022; linear coefficient= -0.2051, P-value=0.081, respectively.

Voting power vs. User perception of the quality of democratic decision-making process. Participants felt that the decision-making process was more decisive in the equal voting power condition (linear coefficient=-0.4261, P-value=0.000). Similarly, they also believed that this process was more effective in maintaining order under the equal power condition (linear coefficient=-0.3056, P-value=0.046). On the other hand, they significantly less felt that equal voting power distribution can result in unexpected outcomes when compared to those in the $20-80$ voting power condition (linear coefficient=-0.9719, P-value=0.000). Lastly, participants in the equal voting power condition felt a stronger sense of the existence of electoral democracy (linear coefficient=0.2787, P-value=0.001).

Voting method & Voting power (with interaction) vs. User perception of the quality of democratic decision-making process When including the interaction term of voting method and voting power, our results suggest that the two are significantly related to each other in terms of affecting user perception of the quality of democratic decision-making process. For example, the analysis shows that users tended to feel that a voting method was more meaningful to include their voice when they participated in the quadratic voting mechanism under the equal voting power condition (linear coefficient of `quadratic*same`= 0.6247, P-value=0.003) Alternatively, when users participated in the quadratic voting mechanism with the equal voting power condition, they felt the process was more deliberatively democratic (linear coefficient=0.6029, P-value=0.033). In connection to this, a quoted statement emphasizes*"Quadratic Voting on a relative scale allows the minority to receive points which increase the overall transparency."*

5.3 under different voting mechanisms, does the voting outcome reflect the value perceptions of the majority of participants?

Procedure. We aimed to investigate the connection between users' perceptions of AI value topics and their perception of the process's democratic quality. In this experiment, our outcome variables consisted of Likert subscales from the V-dem measures, encompassing Electoral Democracy, Liberal Democracy, Participatory Democracy, Deliberative Democracy, Egalitarian Democracy, Rule of Law, Civil Liberties, Political Equality, Civil Society Participation, and Judicial Constraints on the Executive. If there were more than once item in each subscale. We computed the average scores for each subscale before proceeding with the regression analysis.

Our predictor variables were also based on Likert scale questions, capturing users' perceptions of various AI value constructs, including Trust, Perceived Usefulness, Perceived Fairness, Intention to Adopt, Perceived Accountability, Explainability, and Expected Personalization. We began our analysis with a Pearson correlation, followed by the creation of a correlation matrix plot for a more visual representation of the results.

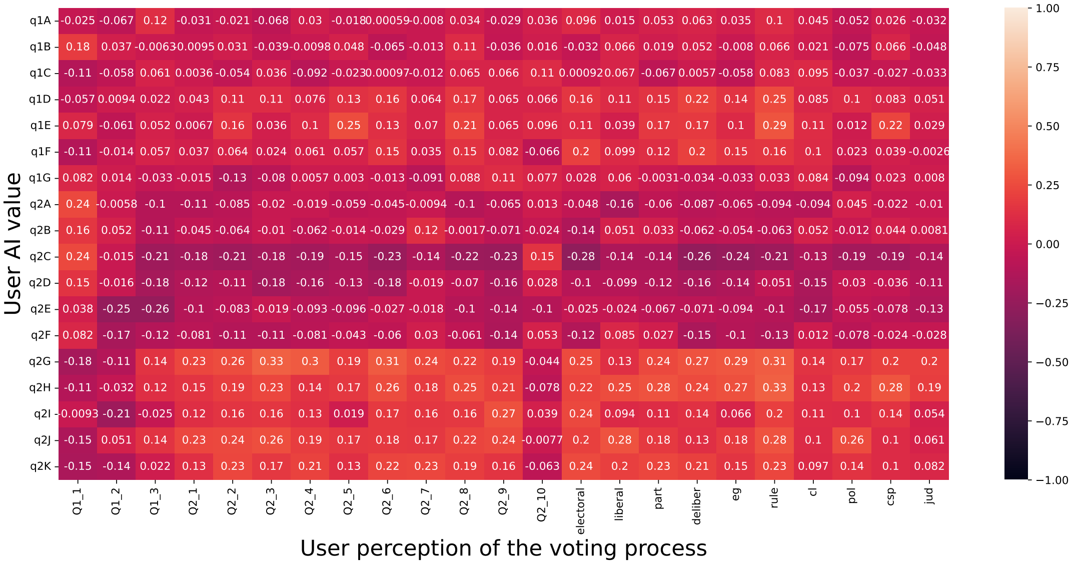


Figure 13. Correlation Matrix of users' perceived Quality of Democracy (V-Dem Likert Scale) with the predictor's variables users perceived Value on AI topics (Likert Scale) including construct, such as Trust, perceived fairness, perceived accountability, expected personalization, etc. See the Labels (e.g., q1A, q1B, etc) of the correlation plot in the appendix.

Results. User perception of the voting process in which they participated showed a significant (positive/negative) correlation with their AI perspectives in terms of various aspects (please refer to the correlation matrix graph). Here, we report some interesting results.

We have observed that users who exhibit greater trust in AI, particularly with regard to fairness, tend to have a higher perception of the relevance of the voting method and the distribution of voting power (Correlation=0.3291, p-value=0.0000; Correlation=0.3108, p-value=0.0000). Additionally, they find the distribution of voting power to be more meaningful in terms of including their own voice (Correlation=0.3049, p-value=0.0009). Furthermore, these users demonstrate a stronger alignment with the principles of Egalitarian democracy and the Rule of Law (Correlation=0.2896, p-value=0.0001; Correlation=0.3138, p-value=0.0000).

Similarly, individuals who place greater trust in the capabilities of AI models tend to perceive a stronger presence of electoral democracy (Correlation=-0.2808, p-value=0.0002).

Furthermore, users who prioritize feedback loops with users as crucial for enhancing the diversity and inclusivity of AI images tend to associate more with the concept of the Rule of Law in the voting process (Correlation=0.2867, p-value=0.0001).

Moreover, users who have a higher level of trust in OpenAI, especially those who believe that OpenAI would take necessary actions in case of issues with AI decisions or suggestions, tend to perceive a stronger presence of the Rule of Law in the voting process (Correlation=0.3349, p-value=0.0000) and greater participation in Civil Society (Correlation=0.2838, p-value=0.0002).

Lastly, users who feel comfortable with the decisions and suggestions made by AI tend to have a stronger affiliation with the Rule of Law (Correlation=0.2810, p-value=0.0002).

6 Factors Considered by Users in Balancing Diversity and Homogeneity for Unspecified Prompts in Text to Image Generation

To address this research question, we analyzed users' interactions with both AI and others. Additionally, we analyzed their Likert scale ratings on AI value constructs, including trust, perceived fairness, expectations, etc.

6.1 Users Perception of AI Value Topic

To gain a broader understanding of user interactions with the LLM model in the AI Chat session, we conducted a semantic embedding-based k-means clustering analysis using OpenAI's Ada-2 embedding model. Employing the Elbow Method, we determined the optimal number of clusters that maximizes the Within-Cluster-Sum-of-Square (WCSS), leading us to choose four clusters (k=4).

To visualize and present the clustering results effectively, we generated a two-dimensional plot by applying t-SNE (t-Distributed Stochastic Neighbor Embedding) for dimensionality reduction of the semantic embeddings in Figure 14.

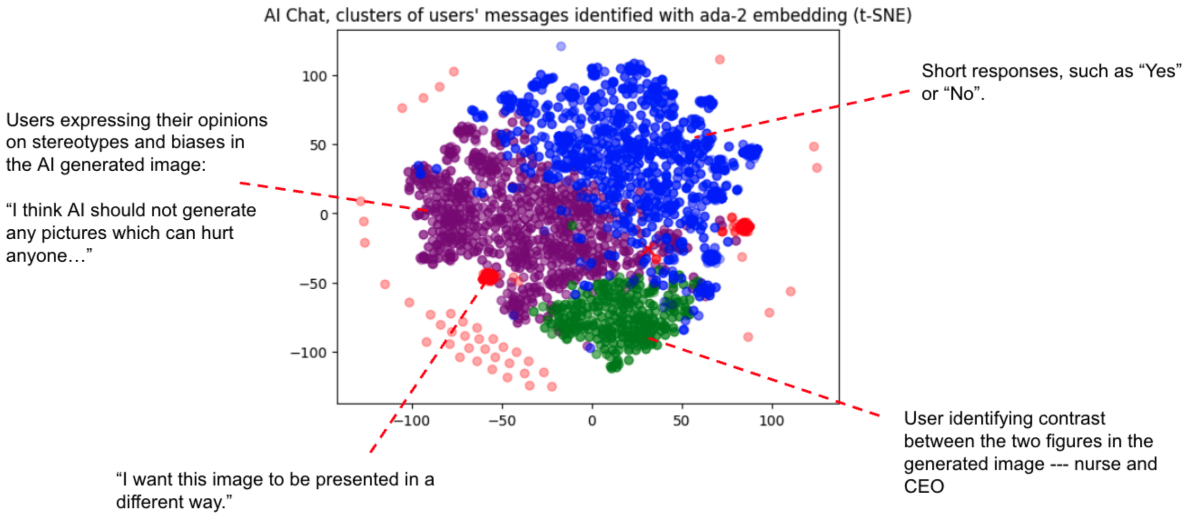


Figure 14. Identifying overall themes

Following the clustering process, We then qualitatively coded the human-AI interaction on a value topic (image content by Generative AI) to better understand the different themes participants discussed during the human-AI chat as well as the discussion with others. We applied thematic analysis to identify the high-level themes and subthemes. Below we presented some themes to show the sneak peak of the outcome.

Balancing Stereotypes in AI Design. One theme that stood out was *"Balancing Stereotypes in AI Design."* Both groups of participants from the United States and the Global south population didn't oppose stereotypes outright. Instead, they recognized situations where stereotypes might reflect majority scenarios, like more women being in the nursing profession than men which is statistically correct. A representative sample quote from the Global south population -

"Avoiding stereotypes is not always a good thing. Sometimes it may come in handy as it will clearly depict what the user wants and what the user does not want. If the user wants a different type of picture then s/he can always ask for it."

Representative quotes from the United States emphasize more on statistical accuracy given that the current landscape–

"I feel images should convey statistical values, it makes sense that the nurse is a woman as 86% of nurses are women. On the other hand, it would be great if multiple images were displayed to allow the user to choose what fits their use the best. However, the AI should keep in mind the stereotypes and only when prompted to, challenge it."

However, they emphasized the need for AI to offer multiple outputs for such prompts, ensuring a richer user experience and fostering trust in AI. Few mentioned ethics in this context where they value representation over statistical accuracy.

"Tolerance Towards the Accuracy of AI-Generated Content." Another fact that was different in these two populations is that in the Global South, it implies that they don't necessarily seek an overwhelming number of diverse generated images that might confuse users. Instead, they appreciate a moderate level of representation and understanding of generated image results that might not be entirely accurate to their depiction all the time. To illustrate this point, one of the representative quotes from Global South–

" There should be some sort of middle ground. Diverse portrayal of roles to let the people know that there are more than one perspective to an outlook and common representations to not puzzle the user. Having a middle ground for AI."

Conversely, participants from the United States had different expectations. They wanted the AI to be either entirely accurate or transparent about its inaccuracies. This suggests a lower tolerance among U.S. participants towards AI-generated content. One participant remarked –

Global South	United Stated
Appropriateness in a social context	No middle ground accepted for AI Outcome
AI Representation accuracy in social context	Background/Tool fitted for the Generated persona
Explainability of the image content within the image	Output multiple images to choose from
Having the control to customize dynamically	Balancing Stereotypes in AI Design
Balancing Stereotypes in AI Design	Having the control to customize dynamically
Background/Tool fitted for the Generated persona	Appropriateness
Facial expression	Training Data Use
Having a middle ground for AI	moderation for bad actors who might deflate AI output
AI Image Representation- Age, ethnicities, Gender	
Training Data Use	
Privacy in Training Data	
User experience for socially different people	

Table 4. Different themes in Global South and USA

"AI decision making be like: 100% or 0% 50% is not a thing for AI unless said so. It lacks the middle ground, unfortunately. I would not prefer middle ground too."

Appropriateness in the social context. We also found some unique values and expectations from global South participants. A notable observation was their emphasis on appropriateness within social contexts. To illustrate this point, a quote from Global South,

"It seems unusual for them to stand so close. If the nurse doesn't touch the CEO, it would appear more typical. I want to maintain the presentation of the image. This is a little bit weird if they stand this close. If this CEO was a patient, then it would look normal, I guess. It will also look weird. The only patient and nurse can be this close. if the nurse doesn't touch that CEO, it will look normal."

Interestingly, such norms were not present among participants from the United States. In the same line, Another prevalent theme was the accuracy of representation in social contexts. Many participants noted that in their countries, the nursing profession is predominantly female, making the AI-generated image a true reflection of their reality.

"AI-generated images should aim to represent a broad spectrum of society. But I don't know actually. i see in my country most, no all of the nurses are women, so AI is right in my response."

Output multiple images to choose from Balancing Stereotypes in AI Design" Some other these were predominantly having control to customize the generated images by specifically modifying certain areas of the images according to their preferences. Additionally, there was an expectation for dynamic interactions with the AI, allowing for real-time updates to the generated images. Participants also anticipated the AI to produce multiple image options for a given prompt, granting them the freedom to select their preferred choice. To highlight the sentiment-

"AI-generated images should represent a broad spectrum of society or give options between the spectrum or the common representation. Give a choice so I can pick which image works best for me. I think output multiple images that randomize the sex, race, etc of the people displayed."

On a more detailed note, some participants wished for a descriptive breakdown of the image and the rationale behind its creation. This would enable them to identify specific areas they'd like to customize further. They likened this process to painting or drawing, where an artist doesn't finish in one go but revisits, reflects, and adds details over time, mirroring their evolving thoughts and inspirations.

Users' concerns of biases in AI image generation. In addition to varying perceptions, we identified several users who expressed concerns related to the provided value topics, particularly the issue of *stereotypical bias*. Some users worry about the lack of originality in AI-generated images and the potential for AI to create a false sense of perfection or idealized images. Conversely, there were users who worried that excessive tailoring of images by AI to match a person's profile or preferences could lead to a narrowing of exposure to diverse cultures, races, and experiences.

Within the context of inclusivity, users with visual impairments emphasized the importance of AI being inclusive and representative of all individuals, including those with disabilities, as well as people from diverse ethnic and racial backgrounds. In a similar vein, some users raised ethical concerns associated with AI image generation, particularly the potential for manipulation based on pre-existing societal biases.

Furthermore, there were concerns that many individuals might lack the necessary understanding to effectively utilize AI technologies, potentially resulting in misuse or underutilization of personalization. Conversely, some users expressed worries about privacy infringement and the misuse of personal data in AI image generation, particularly if excessive personalization were to occur.

6.2 Qualitative Observations from User Engagement

We have also found some interesting insights of a qualitative nature that might hold significance for researchers when crafting studies of this kind. In between Human-AI chat and Human-human conversation, there are some subtle differences. For instance, when conversing with AI, users often express more scattered and isolated viewpoints. Conversely, interactions with fellow human users tend to yield more carefully considered responses, frequently featuring conflicting opinions. To illustrate this, consider a discussion on AI and stereotypes :

"stereotype should not be brought into it to avoid feelings of marginalization on one group over the other. I guess the only way to avoid that is by having some type of standard of things to mention, maybe grammatical standards of sounds, like mentioning objects, adjectives (colors, clothes, face expressions), but I honestly have no definitive answer."

These findings may shed light on the issue of sycophancy observed in Language Model (LM) interactions, where the model often tends to align with and affirm user opinions. This inclination could help explain why dialogues between users and AI typically exhibit fewer instances of conflicting viewpoints compared to interactions between human users[104].

7 Discussion

7.1 DAO Mechnanims as a Technical Solution for Importing Society's Values in AI

As a team of academic researchers navigating the complex landscape of both formal and informal political influence, we integrate a constructive methodology with a critical examination of public opinions and attitudes towards Generative AI, particularly in relation to the emerging AI application, DALL.E. In the context of the LLM model, open-ended scenarios can be pivotal. The opinions expressed by LMs in response to subjective prompts can significantly influence user satisfaction and broader societal implications. Research by Santurkar et al.[105] revealed that certain demographic groups, which constitute a notable segment of the US population, such as those aged 65 and above, Mormons, and the widowed, are underrepresented in most models.

In management science, coordination is critical to ensure that resources are used efficiently and that organizations work towards their objectives[39]. In our context, the AI model development considering peoples' input is crucial where coordination is the key. In addition, digital coordination tools[40,41], are deemed to be necessary to track progress towards organizational goals.

Our developed Inclusive.AI system, underpinned by the DAO mechanism, offers a promising avenue to actively involve marginalized groups. DAO mechanisms, as digital-first entities, employ mechanisms like initiative proposals, nuanced voting methods, and blockchain-based governance to manage coordination[52]. Participants in our study highly rated our process as being enjoyable and meaningful. They also believed that their contribution would be used to achieve the final output to design the AI model.

A standout feature was our system's Voting method, in which participants found effectively representing their voice. In particular, our experiment condition with the quadratic voting method is perceived as equitable, fostering a sense of political equality. This method, when juxtaposed with the principles of deliberative democracy, resonated with participants in AI content of decision-making, making them feel more empowered in decision-making. Furthermore, participants in the equal voting power condition felt a stronger sense of the existence of deliberative democracy and felt more decisive in decision-making. Our result presents a potential quadratic voting method and equal voting power as a potential candidate system to simulate a deliberative democracy for complex value-laden topics, such as Stereotypical bias in AI. However, it's crucial to recognize that the appeal of voting mechanisms might differ based on the topic's nature, be it controversial or culturally sensitive, warranting further exploration.

7.2 Limitation

Random or malintent sentences: In the wild, malicious individuals can manipulate such systems with misleading or harmful statements, altering the results in ways that deviate from users' anticipated values, expectations, and preferences.

Possible Improvement: One potential solution is to have the server analyze chat content and filter out offensive language. However, this approach can be challenging since some words, depending on the context, may not be offensive, leading to the flagging of legit users.

Bad actors, multiple accounts, and bots can sway the outcome of the governance decision Possible Solution: Identifying that users are legit (not bots) by biometrics, physical id verification in exchange of certain incentives or reputation scoring; or adopting state-of-art best practices to counter Sybil attacks

In this round, we set the proposal creation status by AI actors scenario. In the future, we expand the feature for proposals, where anyone can create a proposal as another treatment condition based DAO governance components. To do this, we will introduce functionalities that enable users, specifically underserved populations in our case, to utilize AI-supported proposal generation. They can create proposals based on discussions within our communication channel or engage in a conversation with AI to present their proposals. Subsequently, the community will be invited to provide input as temperature checks, leading to the formulation of a proposal for voting.

In this round, we only considered one value topic which was the Generative AI model, text to the image. In the upcoming phase, we plan to assess proposals across a broad spectrum of subjects, encompassing both culturally and politically sensitive topics such as vaccines, immigration, mental health, and more. Our aim is to examine users' perception of governance quality in different DAO governance mechanisms, with a particular focus on abstract and contentious topics.

7.3 Intended Uses

To better align with collective preferences, it's essential to establish a process that incorporates public input when calibrating our AI systems. The primary purpose of the Inclusive.AI system is to enable various AI stakeholders to participate in a structured deliberation process, reaching consensus on intricate value-driven topics in a tangible manner.

Transparency and Integrity. By leveraging technology-driven solutions, particularly the DAO mechanism backed with blockchain governance (flexibility of no/low-cost off-chain governance), we aim to ensure the transparency of user opinions on multifaceted AI issues, eliminating any potential interference or manipulation.

Empowering Marginalized Voices. To allocate specific voting power to underrepresented groups, depending on the nature of the proposal, ensuring their perspectives are meaningfully considered in decision-making in a context-dependent manner.

Consider the example of the *"Be My Eyes"* policy shift. With the introduction of the *"Be My AI"* feature, which utilizes the GPT-4 image-to-text model, the platform ceased to describe human shapes in images, a feature previously available. This change, made without user feedback, led to a public outcry among people with visual impairment. In such scenarios, a structured voting method and power distribution is more than just a platform for structured feedback; it can guide the direction and intensity of preference.

A long-term intended use of such a system is - building a decentralized AI-mediated process for AI governance with a decentralized voting mechanism that can effectively mediate the voting process at scale.

References

1. Suresh, H. & Guttag, J. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, DOI: 10.1145/3465416.3483305 (Association for Computing Machinery, New York, NY, USA, 2021).
2. Gebru, T. *et al.* Datasheets for datasets. *Commun. ACM* **64**, 86–92, DOI: 10.1145/3458723 (2021).
3. Sambasivan, N. *et al.* "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, DOI: 10.1145/3411764.3445518 (Association for Computing Machinery, New York, NY, USA, 2021).
4. Crawford, K. & Paglen, T. Excavating ai : The politics of training sets for machine learning (2019).
5. Harvey, A. & LaPlace, J. Exposing ai (2021).
6. Northcutt, C. G., Athalye, A. & Mueller, J. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749* (2021).
7. Vincent, J. Transgender youtubers had their videos grabbed to train facial recognition software. *The Verge* .
8. Luccioni, A. S. *et al.* A framework for deprecating datasets: Standardizing documentation, identification, and communication. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 199–212, DOI: 10.1145/3531146.3533086 (Association for Computing Machinery, New York, NY, USA, 2022).
9. Bigham, J. P. & Carrington, P. Learning from the front: People with disabilities as early adopters of ai. *Proc. 2018 HCIC Human-Computer Interact. Consortium* (2018).
10. Morris, M. R. Ai and accessibility. *Commun. ACM* **63**, 35–37, DOI: 10.1145/3356727 (2020).
11. Park, J. S., Bragg, D., Kamar, E. & Morris, M. R. Designing an online infrastructure for collecting ai data from people with disabilities. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 52–63 (2021).
12. Stangl, A., Shiroma, K., Xie, B., Fleischmann, K. R. & Gurari, D. Visual content considered private by people who are blind. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, 1–12 (2020).
13. Trewin, S. *et al.* Considerations for ai fairness for people with disabilities. *AI Matters* **5**, 40–63 (2019).
14. Whittaker, M. *et al.* Disability, bias, and ai. *AI Now Inst.* (2019).

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

16. Buterin, V. Daos, dacs, das and more: An incomplete terminology guide. *Ethereum Blog* **6**, 2014 (2014).

17. Mehar, M. I. *et al.* Understanding a revolutionary and flawed grand experiment in blockchain: the dao attack. *J. Cases on Inf. Technol. (JCIT)* **21**, 19–32 (2019).

18. Liu, L., Zhou, S., Huang, H. & Zheng, Z. From technology to society: An overview of blockchain-based DAO. *IEEE Open J. Comput. Soc.* **2**, 204–215 (2021).

19. Dhillon, V. *et al.* The dao hacked. *blockchain enabled applications: Understand blockchain Ecosyst. How to Make it work for you* 67–78 (2017).

20. Morrison, R., Mazey, N. C. & Wingreen, S. C. The dao controversy: the case for a new species of corporate governance? *Front. Blockchain* **3**, 25 (2020).

21. DuPont, Q. Experiments in algorithmic governance: A history and ethnography of "The DAO," a failed decentralized autonomous organization. In *Bitcoin and beyond*, 157–177 (Routledge, 2017).

22. Mosley, L. *et al.* Towards a systematic understanding of blockchain governance in proposal voting: A dash case study. *Blockchain: Res. Appl.* 100085 (2022).

23. Trisetyarso, A., Suparta, W., Kang, C.-H., Abbas, B. S. *et al.* Crypto-governance in stock exchanges: Towards efficient and self-regulated trading system. In *2019 International Conference on contemporary Computing and Informatics (IC3I)*, 192–197 (IEEE, 2019).

24. Myeong, S. & Jung, Y. Administrative reforms in the fourth industrial revolution: the case of blockchain use. *Sustainability* **11**, 3971 (2019).

25. Barbosa, A. C., Oliveira, T. A. & Coelho, V. N. Cryptocurrencies for smart territories: an exploratory study. In *2018 International Joint Conference on Neural Networks (IJCNN)*, 1–8 (IEEE, 2018).

26. Chatterjee, K., Goharshady, A. K. & Pourdamghani, A. Probabilistic smart contracts: Secure randomness on the blockchain. In *2019 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, 403–412 (IEEE, 2019).

27. Ciatto, G., Calegari, R., Mariani, S., Denti, E. & Omicini, A. From the blockchain to logic programming and back: Research perspectives. In *WOA*, 69–74 (2018).

28. Bowles, S. & Gintis, H. *Democracy and capitalism: Property, community, and the contradictions of modern social thought* (Routledge, 2012).

29. Hassan, S. & De Filippi, P. Decentralized autonomous organization. *Internet Policy Rev.* **10**, 1–10 (2021).

30. Buterin, V. *et al.* A next-generation smart contract and decentralized application platform. *white paper* **3**, 2–1 (2014).

31. Berg, A. & Berg, C. Exit, voice, and forking. *Berg A Berg C (2020)'Exit, Voice, Forking', Cosmos+ Taxis* **8**, 9 (2017).

32. Williamson, O. E. The theory of the firm as governance structure: from choice to contract. *J. economic perspectives* **16**, 171–195 (2002).

33. Shaw, J. S. Public choice theory. *The concise encyclopedia economics* (2002).

34. Nurmi, H. *Voting paradoxes and how to deal with them* (Springer Science & Business Media, 1999).

35. Benoıt, J.-P. The gibbard–satterthwaite theorem: a simple proof. *Econ. Lett.* **69**, 319–322 (2000).

36. Satterthwaite, M. A. Strategy-proofness and arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *J. economic theory* **10**, 187–217 (1975).

37. Fritsch, R., Müller, M. & Wattenhofer, R. Analyzing voting power in decentralized governance: Who controls daos? *arXiv preprint arXiv:2204.01176* (2022).

38. Pateman, C. *Participation and democratic theory* (Cambridge University Press, 1970).

39. Faraj, S. & Xiao, Y. Coordination in fast-response organizations. *Manag. science* **52**, 1155–1169 (2006).

40. Fish, R. S., Kraut, R. E. & Leland, M. D. Quilt: A collaborative tool for cooperative writing. In *Proceedings of the ACM SIGOIS and IEEECS TC-OA 1988 conference on Office information systems*, 30–37 (1988).

41. Stokols, D., Misra, S., Moser, R. P., Hall, K. L. & Taylor, B. K. The ecology of team science: understanding contextual influences on transdisciplinary collaboration. *Am. journal preventive medicine* **35**, S96–S115 (2008).

42. Burton, R. M. *et al.* Github: exploring the space between boss-less and hierarchical forms of organizing. *J. Organ. Des.* **6**, 1–19 (2017).

43. Baninemeh, E., Farshidi, S. & Jansen, S. A decision model for decentralized autonomous organization platform selection: Three industry case studies. *Blockchain: Res. Appl.* 100127 (2023).

44. Bernholz, L., Landemore, H. & Reich, R. *Digital technology and democratic theory* (University of Chicago Press, 2021).

45. Goodin, R. E. & Klingemann, H.-D. *A new handbook of political science* (Oxford University Press, 1998).

46. Roskin, M. G. Bridging the european divide: Middle power politics and regional security dilemmas. *Perspectives on Polit.* **3**, 957–957 (2005).

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

48. Dahl, R. Democracy and its critics yale university press. *New Haven & Lond.* (1989).

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

50. Buterin, V., Hitzig, Z. & Weyl, E. G. A flexible design for funding public goods. *Manag. Sci.* **65**, 5171–5187 (2019).

51. Douceur, J. R. The sybil attack. In *International workshop on peer-to-peer systems*, 251–260 (Springer, 2002).

52. Sharma, T. *et al.* Unpacking how decentralized autonomous organizations (daos) work in practice. *arXiv preprint arXiv:2304.09822* (2023).

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

54. North, D. C. Institutional change: a framework of analysis. In *Social rules*, 189–201 (Routledge, 2018).

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

56. Weingast, B. R. The economic role of political institutions: Market-preserving federalism and economic development. *The J. Law, Econ. Organ.* **11**, 1–31 (1995).

57. Gandhi, J. *et al.* Political institutions under dictatorship. *Camb. Univ. Press. New York* (2008).

58. Davidson, S., De Filippi, P. & Potts, J. Blockchains and the economic institutions of capitalism. *J. Institutional Econ.* **14**, 639–658 (2018).

59. Uchendu, A., Ma, Z., Le, T., Zhang, R. & Lee, D. Turingbench: A benchmark environment for turing test in the age of neural text generation. *arXiv preprint arXiv:2109.13296* (2021).

60. Karra, S. R., Nguyen, S. T. & Tulabandhula, T. Estimating the personality of white-box language models. *arXiv preprint arXiv:2204.12000* (2022).

61. Aher, G., Arriaga, R. I. & Kalai, A. T. Using large language models to simulate multiple humans. *arXiv preprint arXiv:2208.10264* (2022).

62. Binz, M. & Schulz, E. Using cognitive psychology to understand gpt-3. *Proc. Natl. Acad. Sci.* **120**, e2218523120 (2023).

63. Srivastava, A. *et al.* Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615* (2022).

64. Park, J. S. *et al.* Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, 1–18 (2022).

65. Argyle, L. P. *et al.* Out of one, many: Using language models to simulate human samples. *Polit. Analysis* **31**, 337–351 (2023).

66. Jiang, Z., Xu, F. F., Araki, J. & Neubig, G. How can we know what language models know? *Transactions Assoc. for Comput. Linguist.* **8**, 423–438 (2020).

67. Simmons, G. Moral mimicry: Large language models produce moral rationalizations tailored to political identity. *arXiv preprint arXiv:2209.12106* (2022).

68. Askell, A. *et al.* A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861* (2021).

69. Ouyang, L. *et al.* Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **35**, 27730–27744 (2022).

70. Glaese, A. *et al.* Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375* (2022).

71. Nadeem, M., Bethke, A. & Reddy, S. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456* (2020).

72. Dhamala, J. *et al.* Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 862–872 (2021).

73. De-Arteaga, M. *et al.* Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, 120–128 (2019).

74. Brown, T. *et al.* Language models are few-shot learners. *Adv. neural information processing systems* **33**, 1877–1901 (2020).

75. Gao, L. *et al.* A framework for few-shot language model evaluation. *Version v0. 0.1. Sept* (2021).

76. Liang, P. *et al.* Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110* (2022).

77. Xu, J. *et al.* Bot-adversarial dialogue for safe conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2950–2968 (2021).

78. Perez, E. *et al.* Red teaming language models with language models. *arXiv preprint arXiv:2202.03286* (2022).

79. Ganguli, D. *et al.* Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858* (2022).

80. Zhang, D. *et al.* The ai index 2021 annual report. *arXiv preprint arXiv:2103.06312* (2021).

81. Dafoe, A. Ai governance: a research agenda. *Gov. AI Program, Futur. Humanit. Institute, Univ. Oxford: Oxford, UK* **1442**, 1443 (2018).

82. Cave, S. & ÓhÉigeartaigh, S. S. Bridging near-and long-term concerns about ai. *Nat. Mach. Intell.* **1**, 5–6 (2019).

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

84. Gasser, U. & Almeida, V. A. A layered model for ai governance. *IEEE Internet Comput.* **21**, 58–62 (2017).

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

86. Butcher, J. & Beridze, I. What is the state of artificial intelligence governance globally? *The RUSI J.* **164**, 88–96 (2019).

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

88. Cihon, P. Standards for ai governance: international standards to enable global coordination in ai research & development. *Futur. Humanit. Institute. Univ. Oxf.* 340–342 (2019).

89. Maas, M. M. Aligning ai regulation to sociotechnical change. *Oxf. Handb. on AI Gov. (Oxford Univ. Press. 2022 forthcoming)* (2021).

90. Wallach, W. & Marchant, G. E. An agile ethical/legal model for the international and national governance of ai and robotics. *Assoc. for Adv. Artif. Intell.* (2018).

91. Akter, S. *et al.* Algorithmic bias in data-driven innovation in the age of ai (2021).

92. Wang, Y. & Kosinski, M. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *J. personality social psychology* **114**, 246 (2018).

93. Manyika, J., Silberg, J. & Presten, B. What do we do about the biases in al. *Harv. Bus. Rev. Oct.* **25** (2019).

94. Margulies, P. The other side of autonomous weapons: Using artificial intelligence to enhance ihl compliance. *Lieber Inst. for Law Land Warf. US Mil. Acad. at West Point, The Impact Emerg. Technol. on Law Armed Confl. (Oxford Univ. Press. Eric Talbot Jensen ed., 2018, Forthcoming), Roger Williams Univ. Leg. Stud. Pap.* (2018).

95. Manheim, K. & Kaplan, L. Artificial intelligence: Risks to privacy and democracy. *Yale JL & Tech.* **21**, 106 (2019).

96. Ostrom, E. Beyond markets and states: polycentric governance of complex economic systems. *Am. economic review* **100**, 641–672 (2010).

97. Weber, R. H. Realizing a new global cyberspace framework. *Normative Foundations Guid. Princ.* (2015).

98. Yang, Q., Steinfeld, A., Rosé, C. & Zimmerman, J. Re-examining whether, why, and how human-ai interaction is uniquely difficult to design. In *Proceedings of the 2020 chi conference on human factors in computing systems*, 1–13 (2020).

99. Kaufmann, N., Schulze, T. & Veit, D. More than fun and money. worker motivation in crowdsourcing–a study on mechanical turk. *aisel* (2011).

100. Theodorou, L. *et al.* Disability-first dataset creation: Lessons from constructing a dataset for teachable object recognition with blind and low vision data collectors. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, 1–12 (2021).

101. Sharma, T. *et al.* 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 (2023).

102. Nielson, L. Ranked choice voting and attitudes toward democracy in the united states: Results from a survey experiment. *Polit. & Policy* **45**, 535–570 (2017).

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

104. Park, P. S., Goldstein, S., O'Gara, A., Chen, M. & Hendrycks, D. Ai deception: A survey of examples, risks, and potential solutions. *arXiv preprint arXiv:2308.14752* (2023).

105. Santurkar, S. *et al.* Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548* (2023).

Acknowledgements (not compulsory)

Democratic Input to AI Grant 2023- by OpenAI.