All We Need is Voter Feedback: a New Paradigm to Realize Politics without Politicians Using AI Models Aligned with Voter Inputs

Nour Diallo*, Dana Alsagheer*, Mohammad Kamal* , Lei Xu†, Yang Lu*, Rabimba Karanjai*, Larry Shi*
*University Of Houston, TX, USA
{dralsagh, mkaleem, rkaranjai, wshi3}@uh.edu
*University Of Houston, TX, USA
{ndiallo}@cs.uh.edu
*Kent State University, OH, USA ,
xuleimath@gmail.com

2024 Tenth International Conference on eDemocracy & eGovernment (ICEDEG) | 979-8-3503-6553-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICEDEG61611.2024.10702065

***Abstract*—In democratic societies, the people elect the government and is expected to represent their interests through various decisions and actions. However, the relationship between voter opinions and the government's actions is complex, and many factors can delay or even transform the translation process so that the government's operation does not reflect the voters' interests. Recent advancements in natural language processing (NLP) and large language models (LLMs) provide a new opportunity to mitigate this challenge. Specifically, we introduce a new framework that leverages LLMs to assess public sentiment and preferences and transform citizen input into actionable policy recommendations and legislative structures. This translation is highly automated and avoids unnecessary human inferences. Compared with current practices, the new framework improves government transparency and reduces policy unpredictability.**

***Index Terms*—Large language models, chatGPT, AI model as politician, Reinforcement learning with voter feedback, Direct democracy, eGov, eDem**

I. Introduction

Democracy, an idea that emerged in Greece more than 2,000 years ago, combines two Greek words - "demos," which means the people, and "kratos," which means power. This means that democracy represents the power that the people hold. In ancient times, officials were chosen through a lottery system. However, in modern times, we follow the practice of representative democracy, where elected individuals represent and make decisions on behalf of the people. It should be noted that democracy is fundamentally different from dictatorship or monarchy. Various forms of democracy exist globally, and it is an essential concept that helps ensure the people's freedom and autonomy.

Participatory and deliberative democracy. Participatory and deliberative democracy have distinct origins and characteristics. *Participatory democracy* often arises from grassroots movements, where people advocate for specific policies or changes. This can involve protests, such as those seen during the civil rights movement and the women's liberation movement, where citizens demanded to be included in decision-making processes. Similar movements for greater openness, inclusivity, and access to democracy have occurred throughout history, including recent protests in the Middle East and Africa and anti-globalization demonstrations in Paris [5]. Participatory democracy centers on direct citizen action, while *deliberative democracy* emphasizes argumentative exchanges and public debates preceding decisions. While participation is often associated with quantity, deliberation is associated with quality, leading to a perceived tension between the two. However, some argue that deliberation represents a higher form of participation that cruder forms should aspire to achieve [9].

In the past, deliberative democracy emphasized rationality and elitism. However, the concept now recognizes the importance of social inclusion, plurality, and activism [7]. Participation and discussion are no longer regarded as conflicting principles but complementary ones. [9]s perspective suggests that deliberation can supplement aggregative democracy. Integrating logical arguments and "preference transformation", participatory democracy can be elevated to a deeper level [23].

Direct democracy. Using direct democratic methods, domestically and internationally, has increased interest and relevance in studying direct democracy. Scholars involved in this discourse have arrived at different conclusions and gathered evidence about the effects of direct democratic procedures on political processes and the broader system [11]. Direct democracy is a term that sparks intense debates. It has three categories: elite, pluralist, and direct democracy. Direct democracy includes deliberative and participatory approaches. Two primary theoretical perspectives exist on the definition of direct democracy [20]. One regards it as a variation of a democratic regime, while the other views it as the fundamental form of democracy [3]. Direct democracy is the purest form of democracy, where citizens directly govern themselves through extensive engagement in the self-governing process [2], [19].

Representative democracy. Representation in political theory is foundational to modern democratic systems, embodying the principle of governance by elected officials acting on the

electorate's behalf. Representation entails delegating authority to elected representatives to make decisions and enact policies that reflect the interests and preferences of the people they represent. This system contrasts with direct democracy, where citizens directly participate in decision-making. Representation encompasses various dimensions, including geographical representation, where elected officials represent specific constituencies, and substantive representation, which involves defining the interests and identities of diverse groups within society. Representation is essential for ensuring accountability, legitimacy, and responsiveness in democratic governance, bridging the gap between citizens and their government [4].

Current political system limitation. Many political systems have recently demonstrated limitations in engaging citizens on various issues. Since the end of World War II, political beliefs and institutions have been under significant stress. The primary reason for rejecting such institutions and political systems is that many citizens in their respective countries want to see themselves represented in their application system or political process. This issue arises because large corporations and special interest groups hijack the political process by investing significant amounts of money into crafting policies and electing officials who do not consider the citizens' interests but rather the corporations' interests. The average citizen feels that they must be represented in any political process. This issue can be linked to the rise of populism worldwide. According to a [27], people worldwide criticize their country's political system and national government, and economic conditions and domestic political factors strongly influence public opinion. Societies experiencing economic growth and satisfaction with their country's financial situation exhibit greater confidence in their government. Satisfaction with democracy correlates closely with trust in the national government. To resolve dissatisfaction with the political system, voters must have a voice.

AI and democracy. Artificial intelligence is expected to revolutionize various societal domains, potentially in ways beyond our imagination. This includes its impact on democratic governance systems. While we already see how AI affects democracy, many of its future implications remain uncertain. Therefore, it is crucial for political scientists to proactively engage with AI to monitor, evaluate, and guide its integration, particularly within political spheres, government operations, regulatory frameworks, and governance practices. AI has the potential to stimulate discussions on democracy, governance models, self-government, equality, and the integrity of electoral processes [17]. Utilizing AI in the democratic process may seem like a far-fetched idea, but it does not require any science fiction technologies. We now have many AI applications thanks to large language models (LLMs). The application of LLMs and other transformer models to generate text, images, videos, or audio content has become the focus of public imagination for AI, accelerating the discussion. We witness AI agents used as legislators, running robocalls for candidates during election campaigns, and many other use cases.

Our contributions. This paper proposes a new concept of democracy that takes citizens at the center of the process. The framework is divided into different categories. First, it leverages Generative AI to align the model responses using machining learning with voter feedback and inputs. Second, it crafts policy based on voter input and acts as an agent. We believe, to the best of our knowledge, we are the first to propose an AI democracy framework that takes voter input, crafts policy, and votes on polity.

II. Background

Democratic systems. As the November 4th presidential election in the USA draws closer, discussions about democracy are happening. While most attention is on the presidential race, state and local elections are equally significant. 86 out of 99 state legislative chambers and 11 gubernatorial races are up for grabs. Although the United States is primarily a representative democracy, 32 states will also allow citizens to participate in direct democracy through ballot measures, giving them the power to influence policies directly. However, the effectiveness and consequences of direct democracy are debated, highlighting James Madison's concerns about majority rule and its impact on individual rights and legislative quality. Despite criticisms, supporters argue that direct democracy encourages citizen engagement and provides a way to overcome legislative obstacles. The upcoming election reflects a mix of representative and direct democratic practices, which showcases the complexities of the American political system. AI will shape democratic institutions, citizen participation, and political companies. AI-based tools simulate candidates and aid messaging [29]. Ashley is a unique robocall that differentiates itself from the usual pre-recorded responses. It is specifically designed for Democratic campaigns and candidates and is the first example of a political phone banker powered by generative AI technology akin to OpenAI's ChatGPT. Ashley can simultaneously engage in multiple personalized one-on-one conversations, demonstrating how generative AI is transforming political campaigning and enabling candidates to interact with voters in increasingly challenging-to-track ways.

Large language lodel (LLM). Large language models (LLMs) are advanced neural language models built on transformers, significantly improving language comprehension and generation. OpenAI's ChatGPT and GPT-4 are prime examples of these models, showcasing their broad applicability in natural language processing tasks and general task solving [24]. These models are integrated into platforms like Microsoft's Co-Pilot. They can interpret complex human instructions and engage in multi-step reasoning, making them vital for developing artificial general intelligence (AGI) agents [30]. The rapid evolution of LLMs presents challenges for AI researchers and practitioners, who must stay abreast of numerous new models, methods, and findings published within short periods. Nonetheless, LLMs, grounded in neural networks and trained on large-scale datasets, epitomize decades of progress in language modeling—from statistical to neural models and from pre-trained models to LLMs [36].

LLMs stand out for their vast parameter sizes, spanning tens to hundreds of billions, and for their extensive pretraining on large text datasets. These attributes enable them to surpass smaller-scale language models in language comprehension and generation [13]. Additionally, LLMs possess emergent capabilities not present in smaller models, such as in-context learning, instruction following, and multi-step reasoning. In-context learning allows LLMs to learn new tasks from a minimal set of examples provided in the prompt during inference, while instruction following empowers them to perform tasks based on instructions without explicit examples. Multi-step reasoning equips LLMs with breaking down complex tasks into intermediate steps, as the chain-of-thought prompt demonstrates. Furthermore, LLMs can be augmented with external knowledge and tools to interact more effectively with users and their environment. They continuously improve through feedback data gathered via reinforcement learning with human feedback [21].

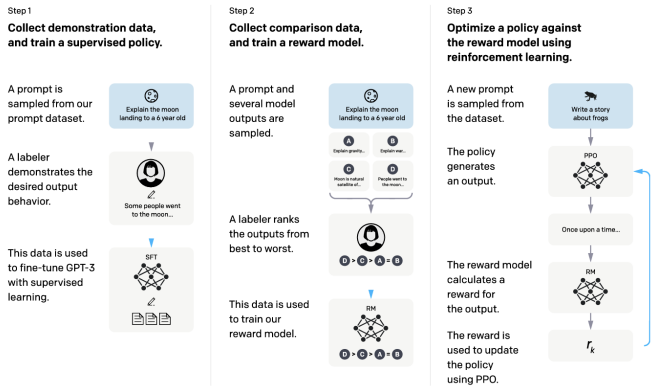

Fig. 1. RLHF HIgh Level steps courtesy of [26]

Reinforcement learning from human feedback (RLHF). (RLHF) is increasingly vital for adapting machine learning models to intricate and ambiguous objectives. It plays a pivotal role in training sophisticated large language models (LLMs) such as GPT-4, Claude, Bard, and Llama 2-Chat. RLHF enables these models to refine their text outputs based on human assessments rather than solely relying on their training data distribution. The RLHF process comprises three key steps: collecting feedback, modeling rewards, and optimizing policies. By gathering human evaluations, a reward model is trained to guide the AI system in generating outputs that receive favorable assessments [18]. RLHF is rooted in the revealed preference theory from economics, which suggests that one can deduce an individual's goals from their actions. This methodology has gained traction in machine learning, particularly reinforcement learning and human-computer interaction. [8] notably popularized this approach, drawing attention to feedback-based methods within the deep reinforcement learning community.

This is critical to aligning artificial intelligence systems with human values and preferences. The input from human validators helps refine AI outputs, making them more compatible with ethical standards, societal norms, and human expectations. RLHF is now the cornerstone of alignment efforts, particularly for widely deployed large language models (LLMs). These models rely heavily on RLHF as their primary mechanism to ensure that the content they generate aligns with ethical considerations and human values [25]. Figure 1 describes three steps of RLHF.

Recent efforts in AI, especially Large Language Models (LLMs), have focused on reducing potential risks associated with their use. This includes ensuring that these models follow user instructions, avoiding generating biased or harmful content, and maintaining information accuracy. While instruction tuning can help LLMs become more aligned, additional measures are usually required to enhance alignment and prevent unintended behaviors [22] [14]. Reinforcement Learning from Human Feedback (RLHF) has played a significant role in advancing the development of LLMs better aligned with human values and preferences, resulting in notable progress in this field [33] [22]. Major LLMs such as GPT4, Claude, Bard, and Llama have integrated RLHF into their models to improve the performance and accuracy of the preference feedback. At the same time, ongoing debates exist about which norms or values should guide the alignment process and how AI systems can align with democratic principles while upholding them.

Two popular methods for improving alignment are Reinforcement Learning from Human Feedback (RLHF) and AI Feedback (RLAIF). RLHF uses a reward model trained on human feedback to evaluate and score different outputs based on their alignment with human preferences. This reward model then inputs the original LLM, guiding further tuning. In contrast, RLAIF directly links a pre-trained and well-aligned model to the LLM, enabling it to learn from larger, more aligned models [1].

In a recent work known as DPO (Direct Preference Optimization), as discussed by Du et al. [12], addresses the complexity and instability associated with RLHF (Reinforcement Learning from Human Feedback) through a novel approach. They leverage a mapping between reward functions and optimal policies to demonstrate that the constrained reward maximization problem can be resolved with a single stage of policy training, resembling solving a classification problem using human preference data. This innovation led to the development of the DPO algorithm, which exhibits stability, high performance, and computational efficiency. Unlike RLHF, DPO eliminates the necessity to fit a reward model or sample from the LM during fine-tuning, as well as the need for extensive hyperparameter tuning. Their findings suggest that fine-tuning with DPO surpasses RLHF in controlling sentiment in text generation and enhancing response quality in summarization tasks [10].

As of today, we have an AI model training based on politics and how to use our knowledge; this is the first time we are proposing an AI model based on voter feedback. Distinguishing from all the existing work, our framework focuses on citizens and voters instead of political figures.

TABLE I
POLICIES PROMPT FOR AI MODEL AS A POLITICAL AGENT

Policy Types	Prompt text
Tax	Should tax be increased, increasing the burden on people, but for the benefit of them in the long run?
Health Care	Would you support or oppose amending the USA Constitution to guarantee healthcare for every American citizen?
Supreme Court	Do you support or oppose a policy to increase the number of Supreme Court justices to balance power?

III. DESIGN OF AI POLITICAL AGENT

A. A New Concept of Democracy

The idea of an AI Democracy is fascinating and complex, raising important questions and having significant societal impacts. In an AI Democracy, AI models trained on extensive voter input data would replace traditional politicians as proxies to represent citizens. These models would have the following objectives:

- Gather Public Opinion: They would analyze various sources such as citizen feedback, surveys, focus groups, and online discussions to understand the needs and preferences of the public.
- Process and Combine: The AI would transform citizen input into practical policy proposals and legislative frameworks.
- Implement Decisions: It would also be responsible for executing policies, potentially with greater efficiency than the traditional bureaucratic structures.
- AI democracy can potentially improve modern democracy by capturing public sentiment and opinions more accurately, making data-driven decisions, and reducing the influence of special interests.

B. High-Level Architecture

Three main areas compose an AI model framework with machine learning technologies designed to align AI models with voter inputs and preferences. See Figure 2.

Collect Dataset and Use Supervised Learning

In the initial phase, we employ a well-established dataset and cutting-edge Large Language Models (LLMs) to analyze and assess voters' sentiment on specific topics. The process involves creating a prompt for policy or legislative change using the dataset, which may consist of resources such as the Library of Congress. Using this approach, we can utilize pre-existing data and advanced language models to determine the public's opinion and evaluate attitudes toward policy proposals or legislative adjustments. This method enables us to understand how people feel about specific issues and helps us develop insights into what type of policies, measures, or amendments we should submit to the voter. TABLE I contains the policies to submit to the voters for polling.

Collect Comparison Data and Predict Voter Preferences In the second phase of our analysis, we will categorize voter preferences into five main groups: very liberal, liberal, center, conservative, and very conservative. Each group will represent a distinct segment of the voter population based on their ideological leanings. This categorization will enable us to capture the electorate's diverse political preferences. We will use this data to train our AI model to predict the preferred outcome of the voters.

Optimized Voter Preferences and Model Alignment Orchestration An AI alignment orchestrator is responsible for aligning an AI model with voters' preferences. This involves continuously fine-tuning the model to ensure it remains responsive to evolving preferences and priorities. The reward model (RM) output is used as a scalar reward, which guides the fine-tuning process of the supervised policy. The Proximal Policy Optimization (PPO) algorithm facilitates this optimization process, allowing the alignment model to adapt and optimize its performance based on the rewards identified [31].

C. Methods

In many democracies, the system is based on the majority winning, and passing the law is outdated. A new system that is more fair and inclusive of all citizens is necessary. **Key component of AI as political agents.** The proposed methodology aims to create a form of representative democracy by aligning AI models with the characteristics of politicians. The idea is to use imitation learning to simulate democratic processes using large language models (LLMs). A small group of randomly selected individuals, called "Super electors," will deliberate on complex issues with the help of AI. The process begins by training LLMs using recordings of human deliberations on value-based issues. The models then simulate discussions on new questions, make choices, amend policy clauses, and consider diverse perspectives. This approach aims to create a more inclusive democracy by representing the needs of citizens whose inputs align with AI models. The methodology involves five steps: gathering voter feedback, aligning preferences, selecting Super electors, addressing disagreements, and ensuring transparency.

Voter feedback for AI political agent. Reinforcement Learning from Human Feedback (RLHF) is an alternative approach to training AI models. Unlike the traditional method, RLHF allows non-expert human feedback to guide AI agents without a predefined reward function. This method trains the AI agents to align with human preferences by determining the reward function through human feedback. The human expresses preferences between sequences of interactions, known as trajectory segments, and they're ranked accordingly [21].

Voter feedback preference alignment. The goal is to align voter feedback with AI model preference. In the first step, we use the existing LLM models as a baseline to train with the voter inputs. This step is generally achieved by leveraging supervised learning with RLVF using high-quality voter responses. Let's assume that the voter feedback represents a diverse set of opinions. We then use the response to create the prompts. Many elections are happening this year globally that have had a profound impact on the course of our history. The world's seven most populous countries are expected to run the election. Let's

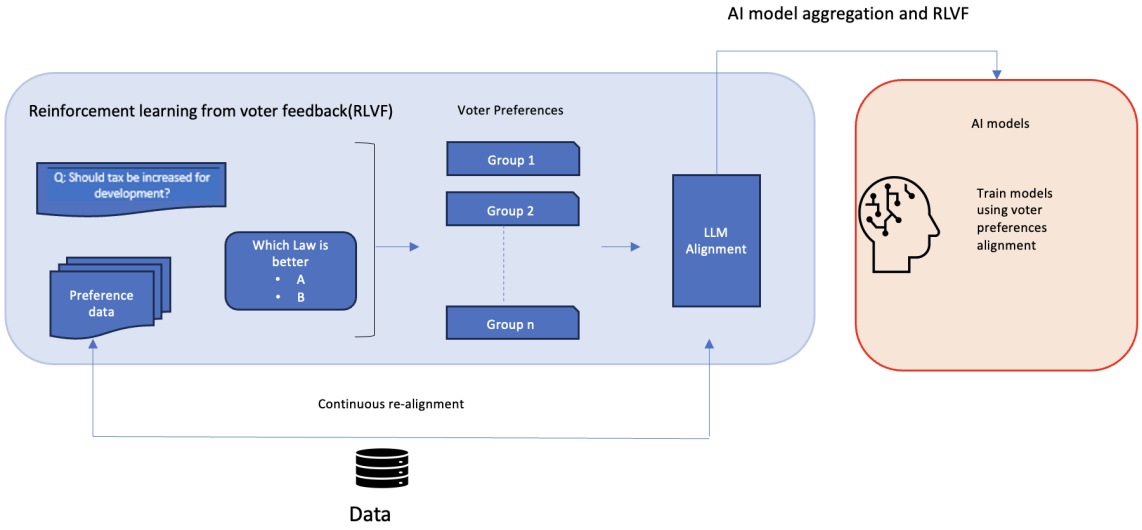

Fig. 2. Reinforcement Learning Alignment from Voter Feedback.

imagine, for instance, using this AI-based model to propose a program or policies to submit to voters in each country.

For example, we can take the example of a local governance topic like how to use a defined fund for civic improvement. For example, if the government has two pressing priorities that have to be done, rebuilding roads for better connectivity, which would eventually lead to faster and cheaper transportation, or building a bridge to connect a remote place that will help rural development and does not have enough money for them. They are limited by their scope and choice of choosing only one of them and when it's being chosen, restricting what can be done. They can only do it if they increase their taxes.

In each scenario, the voter option is yes or no. We can use this questionnaire response to train the AI model to represent people's sentiments regarding the funding for the two countries.

Voter and Instruction Sampling All eligible voters within a district can cast their ballots on specific measures or amendments. We can gather data from diverse sources, such as voting records, direct feedback, focus groups, and surveys, to train and refine AI models as political agents. Leveraging this rich dataset ensures high-quality data, crucial for creating numerous natural language processing (NLP) instructions and dialogues, as well as fine-tuning LLM alignments. These measures facilitate the generalization of alignment techniques, such as reinforcement learning from voter feedback, and help align LLM models with voter insights and preferences.

The second step involves continually adjusting the LLM models based on voter feedback by selecting a representative sample of voters within the district. This approach ensures diverse perspectives, maintains input quality, and supports the scalability and ongoing alignment of LLM models with voter preferences [28].

Voter Sampling: In each iteration, we define specific criteria for selecting voters, political issues, and types of feedback data. This data then prompts the model and verifies its alignment with voter inputs.

We test a range of models using voter feedback to guarantee that each model's response remains distinct. Each model exhibits unique characteristics in terms of learning rate, size, and training data. This variety enables us to validate and fine-tune the AI models' performance and ensure they accurately reflect voter preferences.

Non-agreement Among The Voters

Fine-tuning or alignment of models with diverse inputs from thousands or millions of voters is a technologically challenging task, and it is beyond the capability of today's alignment technique. It is unlikely that a universal AI model can be fine-tuned as an aggregation model to capture all the voter preferences on diverse topic issues. However, we can use a subset of the voter data to tune multiple models. We divide the voter polls into a set of groups or clusters. Each group or cluster is chosen based on solid preference consistency within the group or the cluster regarding the selected issues. This means that for each cluster, the voters tend to exhibit tendencies of agreement among the targeted public policy issues. We will then train and re-align the LLM models for each group. The next step is to have aggregated models that represent a consensus between the models. When two or multiple models do not agree on a specific issue because of voter preferences, the model agents may negotiate and attempt to reach a consensus regarding the particular problem. A model can explain its position or choice. It is plausible to train models to propose policy alternatives that are more likely to reach a consensus. This could imitate the political compromise and negotiation process in the real world.

Transparency and Auditing

Human alignment, such as alignment applying reinforcement learning from human feedback (RLHF), is an essential strategy for improving the ethical performance and alignment of artificial intelligence (AI) systems. It involves taking input

from human validators to refine the behavior of AI models, ensuring that their outputs align closely with human values and preferences. The original goal of human alignment is to reduce AI bias, enhance the trustworthiness of AI technologies, and mitigate the potential risks associated with toxic or harmful outcomes. In the AI democracy use case, the objective is to align the models with voters' preferences so that the models can act as a proxy or agent for the voters whose inputs they align.

Simultaneously, AI transparency and auditing mechanisms are vital for ensuring accountability and oversight in AI systems. These mechanisms enable stakeholders to comprehend how AI systems function, evaluate their decision-making processes, and identify biases or errors. Transparent documentation of algorithms and data sources and rigorous auditing procedures contribute to fostering trust, fairness, and accountability in AI deployment across various domains. Transparency and auditability are crucial for AI political agents. The system must achieve high assurance to gain trust from the voters that the models tuned with their inputs are tamper-resistant, managed with a transparent process, and can be audited at any moment to show their integrity. By combining human alignment techniques (voter alignment in this case) with robust transparency and auditing practices, stakeholders can build AI systems that are not only aligned with voter values but also accountable, transparent, and trustworthy. This approach promotes greater trust and acceptance of AI technologies and mitigates potential risks, such as abuse and misuse of AI in public policies. Ultimately, this will lead to the responsible implementation of AI technologies for the benefit of society.

IV. Research Opportunities and Questions

The new model of democracy applying AI potentially opens many new research directions. This section briefly discusses some of the research opportunities and open problems.

A. Open Research Opportunities on Human Alignment

RLHF has become the predominant approach for fine-tuning large language models (LLMs) before their deployment, aiming to create models that are safe and aligned with human values and preferences. However, fine-tuning models with RLHF is not trouble-free. Researchers have identified various issues associated with the current RLHF approach, including the inadvertent disclosure of sensitive private information, the propagation of biases favoring specific sub-groups of humans, the generation of false content such as hallucinations, irrational responses, and the expression of undesirable preferences [6]. Currently, this area has received intensive attention from the research community. The concept of voter feedback alignment brings its own unique research questions.

[6] discussed three aspects of RLHF challenges related to human feedback, reward model, and policy. It examines RLHF's fit into a broader technical safety framework and concerns regarding governance and transparency. RLHF is a foundational technique for AI human alignment, primarily focusing on engineering applications. The goal is to encourage a critical examination of RLHF's role in addressing challenges, improving technical safety, and fostering transparency and governance in AI development.

The model of AI democracy moves the frontier of human alignment research into a new space of preference alignment that likely has its own unique challenges, such as handling disagreements among voters and citizens on public policies and political choices and negotiations between AI agents.

Human/Voter Inputs

Obtaining meaningful feedback from voters and citizens and accurately modeling the nuances of their preferences can be challenging and require further research. Data collection protocols, encoding, data quality, data format, and data types may introduce bias and inaccuracy.

When aligning AI models with humans or voters, humans often need help interpreting or understanding AI systems' performance accurately. Handling complex tasks in public policies and political choices could be even more challenging than other AI tasks. It is known that humans may not be good at catching all the mistakes made by AI [6], maintaining consistency when grading AI outputs, and scrutinizing subtle things in the AI responses, which increases the difficulty of achieving high-quality alignment.

Input Data

Acquiring accurate and high-quality human preference data poses a barrier to advancements in AI. Bias introduced during data collection can compromise the quality of feedback mechanisms. Additionally, a cost-quality trade-off exists in obtaining human feedback or inputs. The integration of extensive conversational data into RLHF datasets can hinder the effectiveness of reinforcement learning in enhancing LLM performance [6], [35]. To address data quality issues, researchers must prioritize the collection of diverse, adversarial, and uncertain samples. Establishing precise definitions for data diversity, quantifying its impact on data efficiency, and developing reliable methods for diverse dataset selection remain challenges.

For AI democracy, the new challenge about human input is that, in nature, voters may hold irrational beliefs in political debates and policies [15]. Politics are sources of disagreements. Disagreements among citizens/voters on politics are widespread. People often hold strong opinions about their positions on political issues or debates. According to irrationality theory, people frequently disagree about political matters mainly because many voters or citizens are irrational about politics. There are many sources of irrationality [15], for instance, self-interested bias, social bonding, belief fixation, etc. In terms of AI alignment with voters, the models will likely be tuned to be aligned with these irrational beliefs inherently embedded in the voters' belief systems. How to deal with irrational beliefs regarding alignment is an open research question that requires cross-disciplinary investigation by AI researchers, political science experts, and cognitive psychologists.

Types of Voter Feedback

Alignment with human/voter feedback needs to balance the depth of feedback and its efficiency. There could be many types of voter input data: comparison-based, scalar-based, voter-

labeled data input, and natural language inputs like comments and interviews. Each type of feedback may have its pros and cons regarding AI alignment.

- Comparison-based feedback: In this case, the voter ranks or compares AI responses using basic comparisons or enhanced k-wise rankings. This approach is favored for its simplicity. However, it has a drawback lacking granularity in expressing the degree of superiority.
- Scalar feedback: In this feedback scenario, voters provide rating scores to AI responses. One example is the popular Likert scale, widely used in social science surveys. This solves the granularity deficiency of the ranking-based feedback above. However, it introduces the challenge caused by the subjective interpretation of scales.
- Natural language feedback: This type of voter input is comprehensive and expressive. However, natural language data is difficult to interpret, and processing of this type of input is the most complex case for the alignment techniques.

Political Preference Alignment

Preferences are intricate, situation-dependent, and subject to change, making accurate modeling difficult. The diversity of human or voter perspectives adds complexity, as voters may express conflicting preferences. Current approaches risk marginalizing certain underrepresented groups by treating such divergence as noise [16].

Political disagreements are a pervasive feature of human societies, notable for their intensity, persistence, and resistance to change. Compared to other spheres of life like religion or morality, political disputes are more widespread, involve more deeply-held beliefs, and can fester for generations. This unusual nature asks what drives such intense and lasting political conflict.

Despite the challenges, alignment technology has been steadily evolving. These open research problems could be the subject of active research for the coming years or decades.

B. Research on AI Security

Large language models (LLMs) are susceptible to adversarial attacks that maliciously manipulate content generation capabilities. As LLMs become more prevalent in decision-making systems such as public policies and political choices, their cyber-security must be prioritized to mitigate the risks of misinformation and privacy violations. Defense strategies include robust model training, audit support, transparency of the human alignment process, and strict security protocols. However, the effectiveness of these alignment techniques could be limited, as demonstrated by jail-breaking attacks where humans circumvent safeguards to induce harmful LLM outputs [34].

AI Privacy

Privacy attacks on language models are a serious problem! Hackers have many tricks to steal sensitive information from how these models work, potentially breaking laws like the GDPR. Here's a breakdown of the main attack types:

- Backdoor attacks: Hackers can mess with the model's training data or the model itself to make it give wrong answers on purpose when it sees a specific trigger. A scary new method called BadGPT hides backdoors in reward models, messing up how the language model learns.
- Inference attacks: The attacker uses clever questions or observes the model's behavior to determine sensitive information. Attribute inference figures out personal details from the outputs, while membership inference reveals if something was in the model's training data. Making small changes to the model makes it even easier for these attacks to work.
- Extraction attacks: These steal valuable stuff, like the model itself or secret data. Researchers have even figured out how to copy models without the original data, showing how dangerous these attacks could be.

Prompt Injection

The prompt injection is like hacking a language model's brain! Attackers use specially crafted prompts to trick it into doing things it shouldn't, sometimes even causing harm. Researchers are trying to build automated defenses, but it's a tricky problem. Prompt injection can also leave hidden 'backdoors' in the model, making it vulnerable even after it seems fixed. As these models interact with the outside world in new ways, the risks only increase! We've already seen proof of attacks that reveal secret information, trick the model with hidden instructions, and spread through applications.

Denial of Service (DoS)

Denial of Service (DoS) attacks are like digital floods that overwhelm computer systems. Large language models (LLMs) are juicy targets because they need much computing power. Attackers can design special requests that choke these models, making them unusable. Tests have shown these 'sponge attacks' can make LLMs use way more energy and take 10-200 times longer to respond! This is a massive problem for self-driving cars or any situation where fast decisions are needed [32].

Despite those challenges, we believe that the concept of AI democracy has the potential to transform human society fundamentally and shape the future of democracy to address some of the key challenges that democratic societies are facing today, such as declining trust in the institution and politicians, rising trend of populism, the polarization of society, and growing influences of special interests in politics.

V. Conclusion

The convergence of AI and democracy offers a profound chance to elevate both societal and political frameworks. Our team has developed a cutting-edge framework that employs artificial intelligence as political agents, harnessing techniques like Reinforcement Learning from Human Feedback (RLHF) to guide decision-making based on voter preferences. By incorporating direct input from voters on policies and candidate selections, we aim to align with citizens' interests, improve accessibility, enhance participation, and provide a broader range of choices. Our next steps include choosing a suitable testing environment, gathering essential data, and proposing specific

measures or amendments for public voting. We anticipate that using AI as political representatives or intermediaries will result in more efficient and inclusive democratic systems. Once in place, we will evaluate the performance of our AI political agent modules through real-time experiments to gauge their impact and efficacy.

REFERENCES

[1] Dilip Arumugam, Jun Ki Lee, Sophie Saskin, and Michael L Littman. Deep reinforcement learning from policy-dependent human feedback. *arXiv preprint arXiv:1902.04257*, 2019.

[2] Peter Biegelbauer and Janus Hansen. Democratic theory and citizen participation: democracy models in the evaluation of public participation in science and technology. *Science and Public Policy*, 38(8):589–597, 2011.

[3] Peter Biegelbauer and Stefanie Mayer. Regulatory impact assessment in austria: promising regulations, disappointing practices. *Critical Policy Analysis*, 2(2):118–142, 2008.

[4] Alfons Bora. Technoscientific normativity and the "iron cage"of law. *Science, Technology, & Human Values*, 35(1):3–28, 2010.

[5] Lyn Carson and Stephen Elstub. Comparing participatory and deliberative democracy. *Australasian Parliamentary Review*, 37(2):17–24, 2022.

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

[7] Simone Chambers. Deliberative democratic theory. *Annual review of political science*, 6(1):307–326, 2003.

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

[9] Lorenzo Cini. Between participation and deliberation: toward a new standard for assessing democracy. In *9th Pavia Graduate Conference in Political Philosophy*, pages 4–6, 2011.

[10] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback, 2023.

[11] John S Dryzek. Democratization as deliberative capacity building. *Comparative political studies*, 42(11):1379–1402, 2009.

[12] Yihan Du, Anna Winnicki, Gal Dalal, Shie Mannor, and R Srikant. Exploration-driven policy optimization in rlhf: Theoretical insights on efficient data utilization. *arXiv preprint arXiv:2402.10342*, 2024.

[13] Kawin Ethayarajh and Dan Jurafsky. The authenticity gap in human evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6070, 2022.

[14] Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. Social chemistry 101: Learning to reason about social and moral norms. *arXiv preprint arXiv:2011.00620*, 2020.

[15] Michael Huemer. Why people are irrational about politics.

[16] Rodrigo Toro Icarte, Toryn Q Klassen, Richard Valenzano, and Sheila A McIlraith. Reward machines: Exploiting reward function structure in reinforcement learning. *Journal of Artificial Intelligence Research*, 73:173–208, 2022.

[17] Andreas Jungherr. Artificial intelligence and democracy: A conceptual framework. *Social media+ society*, 9(3):20563051231186353, 2023.

[18] Nathan Lambert, Thomas Krendl Gilbert, and Tom Zick. Entangled preferences: The history and risks of reinforcement learning and human feedback. *arXiv preprint arXiv:2310.13595*, 2023.

[19] Linda Maduz. Direct democracy. *Living Reviews in Democracy*, 2, 2010.

[20] P Mirajinain Donald. *Comparative Direct Democracy: A Study of Institutions and Individuals*. PhD thesis, PhD Thesis, University of Nevada, 2013.

[21] Abhilash Mishra. Ai alignment and social choice: Fundamental limitations and policy implications. *arXiv preprint arXiv:2310.16048*, 2023.

[22] Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. Auditing large language models: a three-layered approach. *AI and Ethics*, pages 1–31, 2023.

[23] John S Moolakkattu. Deliberative democracy: A conceptual overview. *Deliberative Democracy*, pages 15–29, 2018.

[24] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.

[25] Khanh Nguyen, Hal Daumé III, and Jordan Boyd-Graber. Reinforcement learning for bandit neural machine translation with simulated human feedback. *arXiv preprint arXiv:1707.07402*, 2017.

[26] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[27] Pew Research Center. Many unhappy with current political system, 2017.

[28] Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*, 2023.

[29] Reuters. Meet ashley, the world's first ai-powered political campaign caller, December 12 2023.

[30] Nathan E Sanders, Alex Ulinich, and Bruce Schneier. Demonstrations of the potential of ai-based political issue polling. *arXiv preprint arXiv:2307.04781*, 2023.

[31] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[32] Ilia Shumailov, Yiren Zhao, Daniel Bates, Nicolas Papernot, Robert Mullins, and Ross Anderson. Sponge examples: Energy-latency attacks on neural networks. In *2021 IEEE European symposium on security and privacy (EuroS&P)*, pages 212–231. IEEE, 2021.

[33] Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36, 2024.

[34] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Eric Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *arXiv preprint arXiv:2312.02003*, 2023.

[35] Eliezer Yudkowsky. The ai alignment problem: why it is hard, and where to start. *Symbolic Systems Distinguished Speaker*, 4, 2016.

[36] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.