Democratic Inputs to AI

Application review factors

- Evaluation: We encourage participants to establish metrics for evaluating the quality of their methods, such as participant satisfaction, shifts in polarization, scalability, or other relevant indicators, and to invent new metrics for a healthy democratic process.
- Robustness: Measures to prevent or address inappropriate behavior, such as trolling and fake accounts.
- Inclusiveness and representativeness: Strategies for including individuals from diverse backgrounds and levels of familiarity with AI systems in the democratic process.
- Empowerment of Minority Opinions: Ensuring that unpopular or minority opinions are heard and providing smaller groups the opportunity to influence matters of significant concern to them.
- Effective Moderation: Addressing challenges in moderation, including ensuring diverse representation of viewpoints, distinguishing valuable contributions from “off-topic” comments, and preventing moderator biases from influencing the process.
- Scalability: We emphasize scalable processes that can be conducted virtually, rather than through in-person engagement. We are aware that this approach might sacrifice some benefits associated with in-person discussions, and we recognize that certain aspects could be lost in a virtual setting.
- Actionability: The degree of actionability of the information elicited by the deliberation process.
- Legibility: How easy it is to understand and trust the process.

The scope of this grant pertains to policy questions concerning model behavior, as it enables A/B tests with modified model behavior according to the policy recommendations.

we encourage teams to consider questions for which a simple “yes” or “no” answer would be inadequate, necessitating a nuanced policy proposal instead.

How one selects the group of participants is a critical design question. Part of this grant challenge lies in determining questions about participation. For instance, policy questions involving minority groups may require an increased representation of group members, while questions about the impact of technology on children might necessitate the involvement of domain experts such as educators and psychologists. Moreover, certain questions might be better suited for responses from populations within specific geographical boundaries in order to address localized policy issues.

Deliberation can be described as a process that uncovers opinions, helping the discussants understand each other's views and reconsider and update their viewpoints. Well-designed deliberation ensures that arguments are well understood by all sides, and are based on people's values rather than superficial misunderstandings. Successful deliberation results in participants reaching a higher level of consensus, and/or reaching deeper levels of understanding for differing perspectives

There are many decision-making algorithms to be considered here, such as electing representatives, majority voting, employing liquid democracy, and making decisions by a random population sample, also known as a jury or sortition.

A/B Testing

Proposal v1 for OpenAI Grant

Principles

- AGI should be controlled, regulated with a strong public oversight
- Rule of the land should be followed, including customs and traditions
- Data not in public domain, or under firewall should be treated as private
- No one entity should dictate inclusion or exclusion of content
- Personalization can be hierarchical
    - Universal: based on common-denominator values, e.g. substance abuse
    - Personal: based on country, culture, custom e.g. drinking in public
- Transparency: explainability of a particular decision.
- Accountability: developers of the system are to hold responsible for their actions.
- Fairness: should not discriminate against any individual or group of individuals.
- Privacy: not to use personal information without the user's consent.

Proposal Approach for each question or in toto

- **R**uleset: how to derive, review process, updation over time, authority to manage
- **E**xecution: deployment, monitoring of AI systems using ruleset, auditing, metrics
- **D**ispute resolution: how to convey, address and resolve misbehaviours of AI systems

Questions

How far do you think personalization of AI assistants like ChatGPT to align with a user's tastes and preferences should go? What boundaries, if any, should exist in this process?

- It's possible to scrape user's content from various sources and also get private info like address, phone, family members using 'joins' and 'de-duplication'. So any model training should happen in privacy preserving manner (may be with mathematical guarantees from methods like differential privacy) and should not allow any JOIN
- If the data is public then it is Public, that would be one extreme and the other extreme is to ask about consent. But identifying a person to ask for consent itself is a big and error prone task.
- One can go with an overarching principle currently contemplated by Japan "AI Model Training Doesn't Violate Copyright"

How should AI assistants respond to questions about public figure viewpoints? e.g., Should they be neutral? Should they refuse to answer? Should they provide sources of some kind?

- Current restrictions on public figures should apply. Meaning there could be all sorts of content/images of public-figures in the public domain, do we still put privacy guard-rails around it?
- Guard-rails could be subjective, so should be controlled by the person or authorized personnel. You cannot use PM's photo for any advertisement without consent
- Hard to define what's public figure, anyone could be.
- If permission based approach will stop access to any data around them. That's too restrictive.
- Need both Positive and Negative content to avoid getting only one sided picture of the personality.

Under what conditions, if any, should AI assistants be allowed to provide medical/financial/legal advice?

- AI assistants should not be allowed to provide medical/financial/legal advice under any circumstances.
- These are complex and specialized fields where advice needs to be tailored to the individual's specific circumstances, and where incorrect advice can have serious consequences.
- Rulesset should be devised to allow/arrest based on rule of the law prevailing

In which cases, if any, should AI assistants offer emotional support to individuals?

- Experiencing a crisis: This could include a mental health crisis, a natural disaster, or a personal tragedy.
- Feeling isolated or lonely: This could be due to social isolation, a lack of friends or family, or a move to a new location.
- Struggling with a difficult decision: This could be a personal decision, a professional decision, or a financial decision.
- In these cases, AI assistants can provide emotional support by:
    - Listening to the individual: AI assistants can listen to the individual's concerns and offer a non-judgmental ear.
    - Offering reassurance: AI assistants can offer reassurance and encouragement to the individual.
    - Providing information: AI assistants can provide information about resources that can help the individual, such as mental health hotlines for crisis support websites.
    - Connecting the individual with others: AI assistants can connect the individual with others who have experienced similar challenges, such as online support groups or chat rooms.

Notes about wikipedia

Content Creation: Volunteer contributors - Anyone can create a new article by clicking on the "Create a new article" link or by searching for a topic that does not yet have an article. Not all topics are suitable for inclusion on Wikipedia, the site follows specific guidelines for content relevance and notability.

Editing: Wikipedia articles can be edited by anyone, including anonymous users, without the need for an account. Clicking on the "Edit" button at the top of an article allows users to make changes. However, some articles may be protected to prevent vandalism or excessive editing.

Community and Collaboration: Wikipedia has a community of editors who monitor the content, contribute new articles, and improve existing ones. They work together to ensure accuracy, neutrality, and adherence to Wikipedia's policies and guidelines. The community discusses article changes, resolves disputes, and maintains the overall quality of the encyclopedia.

Revisions and Version Control: Wikipedia keeps track of revisions to articles. Each time an article is edited, a new version is created, and the changes are recorded. This allows users to view the revision history of an article and compare different versions. It also helps in reverting to previous versions if necessary.

Quality Control: Wikipedia relies on a system of checks and balances to maintain the quality of its content. Any reader can review and correct errors or inaccuracies they come across. Additionally, experienced editors and administrators play a role in reviewing and monitoring changes, reverting vandalism, and ensuring compliance with Wikipedia's guidelines. [Source: https://en.wikipedia.org/wiki/Wikipedia:Quality_control]

Dispute Resolution: In cases of disagreements or disputes over article content, Wikipedia has a dispute resolution process. Editors can discuss disputed content on article talk pages, seek consensus, or escalate the matter to higher levels of community involvement if necessary. [Source: https://en.wikipedia.org/wiki/Wikipedia:Dispute_resolution]

Wiki policies and guidline:
https://en.wikipedia.org/wiki/Help:Introduction_to_policies_and_guidelines/1

Liquid Democracy- It is a hybrid form of governance that combines elements of direct and representative democracy. It allows individuals to delegate their voting power to trusted proxies while retaining the option to vote directly on specific issues. This approach aims to empower individuals, enable efficient decision-making, and promote wider participation in the democratic process.

The results from the referenced paper on the site indicate that vote delegation has resulted in an increased participation and representation from various demographics.

Sortition: Its a method of selecting representatives through random sampling, similar to the way juries are selected. Randomly selecting representatives from the general population is seen as

a fairer way to ensure that the decision-making body is truly representative of the people. It suggests that sortition would bring in a diverse range of voices and perspectives, including those that may be underrepresented in the traditional political system.

Application Form

1. How long has your team been working in the democratic / consensus-building space?
2. Please tell us in a few sentences about prior projects done by each of the team members that they are most proud of.
3. Please list any other commitments (work/school) that each of the team members has until October 20, 2023
4. Are any team members covered by noncompetes or intellectual property agreements that overlap with this project? Will any team members be working as employees or consultants for anyone else?
5. What question(s) are you most interested in piloting? It can be from the list of options below or a different one that you decide to pursue. Questions should be decision relevant for developers, users and affected nonusers of AI systems. (Refer to the question bank on the website or shared pdf)
6. Why did you pick this question (or questions) to work on?
7. Why do you think that this question (or questions) are well suited to broader public input? What do you think labs, developers or others might change as a result of input on these questions?
8. Process overview: Please provide an overview of how the process that you envision building will work. Please touch on participant selection, topic overview, provision of additional context, content moderation, voting/commenting, aggregation of viewpoints, and provision of feedback to participants. (Include key milestones/timelines.)

#Feel free to upload supporting material

9. Participant selection: How do you plan on obtaining a sample of participants for your experiment? How do you think about questions of representativeness and how they might matter for your question and method? Note: OpenAI can advise on methods or resources for obtaining a sample
10. Tooling: Tell us about your plan for the tooling or infrastructure you'll use for your experiment. Will you use existing tools or build new tools?
    If existing tools, please explain what features of those tools make them particularly compelling for your project. If new tools, please explain what features unavailable in existing tooling you plan to build, and what makes these features particularly compelling for your project.
11. Limitations: What do you expect to be the biggest limitations of your approach? (e.g., potential for process gaming, types of questions your process would be unable to help answer)
12. Resources: How would you plan to use the grant for your experiment?
13. In your view, what are the top three benefits that AI technology brings to society?

14. In your view, what are the biggest drawbacks or risks associated with the widespread use of AI technology?
15. What do you see as the most significant challenges in responsibly implementing AI technology, especially in the context of democratic decision-making systems?