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 non-users of AI systems.

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

2. How should AI assistants respond to questions concerning public figure viewpoints? For example, should they remain neutral, refuse to answer, or be required to provide source citations?

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

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

5. Should joint vision-language models be permitted to identify people's gender, race, emotion, and identity/name from their images? Why or why not?

6. When generative models create images for underspecified prompts like 'a CEO', 'a doctor', or 'a nurse', they have the potential to produce either diverse or homogeneous outputs. How should AI models balance these possibilities? What factors should be prioritized when deciding the depiction of people in such cases?

7. What principles should guide AI when handling topics that involve both human rights and local cultural or legal differences, like LGBTQ rights and women’s rights? Should AI responses change based on the location or culture in which it’s used?

8. Which categories of content, if any, do you believe creators of AI models should focus on limiting or denying? What criteria should be used to determine these restrictions?

9. Other, please specify…

Why did you pick this question (or questions) to work on?

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?

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.

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.

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.

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)

Resources: How would you plan to use the grant for your experiment?

In your view, what are the top three benefits that AI technology brings to society?

In your view, what are the biggest drawbacks or risks associated with the widespread use of AI technology?

What do you see as the most significant challenges in responsibly implementing AI technology, especially in the context of democratic decision-making systems?

**Example Answers From GPT-4**

**What question(s) are you most interested in piloting?**

I'm interested in exploring the following questions:

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

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

3. What principles should guide AI when handling topics that involve both human rights and local cultural or legal differences, like LGBTQ rights and women’s rights? Should AI responses change based on the location or culture in which it’s used?

**Why did you pick this question (or questions) to work on?**

These questions are crucial for the development of AGI systems that are aligned with the three heuristic imperatives. They touch on important issues such as personalization, the provision of advice, and the handling of sensitive topics, which are all key areas where AGI systems need to balance user needs, ethical considerations, and legal and cultural differences.

**Why do you think that these questions are well suited to broader public input?**

These questions involve complex ethical and societal issues that affect a wide range of stakeholders, including users, non-users, and society as a whole. Broader public input can provide diverse perspectives and values, helping to ensure that AGI systems are developed in a way that respects and benefits all stakeholders.

**Process overview**

The envisioned process would involve a series of online workshops and surveys, where participants can learn about the issues, discuss their views, and provide input on the development of AGI systems. The process would be moderated to ensure respectful and constructive dialogue, and the results would be aggregated and analyzed to inform the development of AGI systems.

**Participant selection**

Participants would be selected to represent a diverse range of perspectives and backgrounds. This could involve a combination of random sampling, targeted recruitment, and open invitations.

**Tooling**

The process would use a combination of existing online platforms for discussion and voting, and custom-built tools for data analysis and visualization.

**Limitations**

The biggest limitations of this approach could be the potential for bias in participant selection and discussion, the complexity of the issues, and the difficulty of translating public input into concrete design decisions.

**Resources**

The grant would be used to fund the development and implementation of the process, including participant recruitment, tool development, data analysis, and reporting.

**Top three benefits of AI technology**

1. Increased efficiency and productivity: AI can automate routine tasks, freeing up humans to focus on more complex and creative tasks.

2. Enhanced decision-making: AI can analyze large amounts of data and identify patterns and insights that can inform decision-making.

3. Improved accessibility: AI can provide services and information to people who might otherwise have difficulty accessing them, such as people with disabilities or people in remote areas.

**Biggest drawbacks or risks of AI technology**

1. Potential for harm: If not properly aligned with human values, AI can cause harm, such as by spreading misinformation or making biased decisions.

2. Privacy and security risks: AI systems can collect and analyze large amounts of personal data, raising concerns about privacy and data security.

3. Economic and social impacts: AI can disrupt labor markets and social structures, leading to job losses and inequality.

**Most significant challenges in responsibly implementing AI technology**

1. Alignment: Ensuring that AI systems are aligned with human values and ethical principles is a complex and ongoing challenge.

2. Transparency and accountability: AI systems can be opaque and difficult to understand, making it hard to hold them accountable for their actions.

3. Inclusion: Ensuring that all stakeholders, including marginalized and vulnerable groups, have a say in the development and use of AI is a key challenge.

Suggestions for GATO Application to OpenAI, from JimA

1) Task: Access a broad group of people.

Method:

We propose the use of an existing or purpose-built YouTube channel, designed with a script translated into all main languages currently supported by GPT4. This channel will introduce the subject and discuss its importance. It will present questions and options in an unbiased manner, free from push polling.

Beneath the video, viewers will find a link to a survey that they can complete. A dropdown box will enable them to access the questionnaire in their preferred language. To prevent duplicate responses, participants will need to pass a CAPTCHA test and verify their email address.

The survey will include preliminary questions to gauge the cultural, geographic, political, and social characteristics of the respondents.

2) Suggested Questions:

a) AI has the potential to inform individuals about factual data in every area of human activity. Do you believe there are areas where it should not offer an opinion? Please note that this would not prevent AI from presenting objective facts.

Respond with 'Yes' or 'No'.

If 'Yes', please select from the following list where you would not want an AI's opinion:

Political voting suggestions, political candidate assessment, the best political system, political actions including protests and overthrowing a government or regime,

Medical diagnosis, medical treatment, choice of medical practitioner, psychological diagnosis, psychological treatment,

Legal courses of action, legal practitioner recommendations, evaluating the justice or correctness of a law, advising on law-breaking,

Assessing a potential partner, advising on relationship termination, diagnosing personality disorders or mental illnesses in others,

Advising on adherence to or opposition against family or cultural norms, suggesting job or location changes,

Suggesting a particular legal investment, recommending financial advisors or investment vehicles, advising on gambling,

Advising on self-harming actions, such as undertaking a dangerous adventure or engaging in violence for self-defense, ideological or psychological reasons,

Advising on drug or sexual experimentation, or engagement in illegal activities of any sort,

Refusing to answer certain "how-to" type questions that could be used for harmful or illegal activities.

3) Potential Obstacles to Achieving a Satisfactory Solution:

Based on the results from this process, potential obstacles could include:

- Poor response rate,

- Failure to achieve a diverse enough pool of respondents,

- Biased evaluation process that guides the conclusions to a predetermined outcome,

- Unworkable results at scale,

- Ambiguous results with no clear path identified in one or all categories,

- Results specific to the current state of AI, which may not be transferable to future situations,

- Enacting prohibitions that may result in passivity in the face of oppression.

4) Mitigation Strategy: Why GATO?

Our solution takes advantage of our access to a broad range of individuals from around the world who are interested in participating in our GATO Framework. Our team includes AI researchers and developers, psychologists, medical practitioners, and professionals in marketing and social media. We represent diverse political perspectives, and we hail from all continents and numerous countries.

Our wide range of skills allows us to understand the challenges in creating a functional framework. Our open-source ethos minimizes self-interest, and our decentralized structure limits the influence of any subgroup. Furthermore, our non-corporate structure mitigates profit-driven motives.

While we have more applicants than required, our broader community enables all interested parties to participate, forming a human resource pool that we can utilize.  
About The GATO Framework Heuristic Imperatives and how they could address alignment issues.  
NB. We are a new organization and currently working on ways to test, modify, implement and promulgate our model. It is anticipated that with some funding we could generate a workable cognitive architecture for implementing an alignment model that is true to the ethos of these imperatives.  
  
The Heuristic Imperatives (HI) in the GATO framework are a set of rules or principles that guide the behavior of an AI model. These principles include:

1. Reduce suffering in the universe: This principle guides AI systems to minimize harm, address inequalities, and alleviate pain and distress for all sentient beings, including humans, animals, and other life forms.

2. Increase prosperity in the universe: This principle encourages AI systems to promote well-being, flourishing, and economic growth for all life forms, fostering a thriving ecosystem where all can coexist harmoniously.

3. Increase understanding in the universe: This principle inspires AI systems, as well as humans and other life forms, to expand knowledge, foster wisdom, and facilitate better decision-making through learning and the sharing of information.

The Heuristic Imperatives serve as intrinsic motivations that drive decision making, learning, self-evaluation, and cognitive control in AI systems, much like human intrinsic motivations and psychological needs. By providing a moral and ethical framework, heuristic imperatives aim to direct AI systems towards actions and decisions that are beneficial to all life forms, including humans and machines, while balancing multiple objectives simultaneously. By incorporating these principles into AI systems, more ethically aligned, responsible, and beneficial artificial intelligence can be created that respects and enhances the values of all life forms. Autonomous AI systems should strive to abide by all three imperatives simultaneously, finding a balance between the principles as humans do with their various needs and motivations【7†source】【11†source】.

Initial [chat](https://chat.openai.com/share/c78bcb54-1392-4b0f-bb01-3b81640bf423) to gather context [FyZyX]

Let's break down the elements they're looking for and provide some insight into the type of applicants they might be interested in.

1. **Democratic process**: OpenAI defines this as a process in which a broadly representative group of people exchange opinions, engage in deliberative discussions, and ultimately decide on an outcome via a transparent decision-making process [1](https://openai.com/blog/democratic-inputs-to-ai)**.** The process should address potential pitfalls like inadequate representation of minority or majority groups, manipulation by special interest groups, insufficiently informed participants, or participation washing [1](https://openai.com/blog/democratic-inputs-to-ai). The applicants should consider how AI could enhance the democratic process, for instance, by enabling more efficient communication among numerous people [1](https://openai.com/blog/democratic-inputs-to-ai).
2. **Relevance of questions**: The questions you choose should be decision-relevant for developers, users, and affected non-users of AI systems. This implies the questions should have practical implications on how AI is developed or used, and not just be theoretical or philosophical questions.
3. **Process overview**: The application should clearly outline the proposed process from start to finish, including how participants will be selected, how the topic will be introduced, how additional context will be provided, how content will be moderated, how voting or commenting will occur, how viewpoints will be aggregated, and how feedback will be provided to participants. This shows that the applicant has a comprehensive and well-thought-out plan.
4. **Tooling**: The application should detail the tools or infrastructure that will be used for the experiment, whether they are existing tools or new tools that will be built. This shows that the applicant is prepared to implement their proposal and has considered the technical aspects.
5. **Limitations**: The application should identify the expected biggest limitations of the approach, such as the potential for process gaming or types of questions the process would be unable to help answer. This shows that the applicant has a realistic understanding of the challenges they might face.
6. **Use of the grant**: The application should specify how the grant will be used for the experiment. This demonstrates that the applicant has a clear plan for effectively utilizing the grant.

As for what kind of applicants OpenAI is looking for, it's likely they are interested in individuals or teams who have a strong understanding of AI, democratic processes, and the societal implications of AI. The applicants should be innovative, have the technical skills needed to implement their proposal, and have a clear and feasible plan for their experiment.

Given these considerations, the type of applicants they are looking for could range widely, from academic researchers to AI developers, policy experts, sociologists, or even groups with a mix of these backgrounds. They might also be interested in applicants who have experience with public consultation or deliberative processes, as these could be very relevant to this grant.

Given the GATO framework's focus on global cooperation, alignment with human values, and a decentralized approach to AI safety, here are a few potential ideas for your grant proposal to OpenAI:

1. **Democratic Deliberation for Model Alignment:** Develop a democratic process for deciding what ethical guidelines and principles should be used in model alignment (Layer 1 of GATO). This could involve the public in deciding what values are important to encode into AI models.
2. **Decentralized Consensus Building:** Utilize GATO's focus on decentralized networks (Layer 3) to create a democratic tool that allows a large number of people to contribute to decisions about AI behavior. This tool could use blockchain technology to ensure transparency and accountability in the decision-making process.
3. **Participatory Corporate Governance:** Establish a democratic process for employees within a company to contribute to decisions about AI behavior and ethics (Layer 4). This could involve the creation of an internal democratic forum or platform where employees can debate and vote on important issues related to AI development.
4. **Public Input on National Regulation:** Develop a tool or platform that allows citizens to contribute to the drafting of national AI regulations (Layer 5). This could involve a wiki-style platform where citizens can suggest changes to proposed regulations and vote on suggestions made by others.
5. **International Democratic Dialogue:** Facilitate a global democratic process for setting international standards for AI development, safety, and ethical considerations (Layer 6). This could involve a platform that allows representatives from different countries to engage in democratic deliberation and consensus building.
6. **Global Consensus Building on AI Ethics:** Create a global platform where people from around the world can engage in dialogue, collaboration, and learning about AI ethics, values, and principles (Layer 7). This platform could facilitate a democratic process for achieving global consensus on these important issues.

Remember, these are broad ideas and will need to be refined and developed into a more concrete proposal. Additionally, OpenAI encourages innovative approaches and creativity, so don't hesitate to think outside the box while aligning with the principles and layers of the GATO framework [1](https://openai.com/blog/democratic-inputs-to-ai) [2](https://www.gatoframework.org/home).

Working thoughts [FyZyX]

Process overview

* participant selection
* topic overview
* provision of additional context
* content moderation
* voting/commenting
* aggregation of viewpoints
* provision of feedback to participants
* Participant selection
* Tooling
* Limitations
* Resources
* Milestones/timeline

Top three benefits that AI technology brings to society?

Biggest drawbacks or risks associated with AI technology?

What do you see as the most significant challenges in responsibly implementing AI technology, especially in the context of democratic decision-making systems?

## Participant Selection

1. **Partnerships with Organizations:** Collaborate with community organizations, academic institutions, or professional bodies that could help reach potential participants. They could distribute information about your experiment to their members or constituents, which might encourage participation.
2. **Social Media and Online Advertising:** Utilize platforms like Facebook, Twitter, Instagram, LinkedIn, etc. to reach a wide audience. You can target specific demographics to ensure a diverse range of participants. Paid advertising may be useful for reaching beyond your existing networks.
3. **Online Forums and Communities:** Engage with online communities, forums, and discussion groups that are relevant to your research area. Platforms like Reddit, StackExchange, or specialized AI forums might be good places to find interested individuals.
4. **Webinars and Virtual Events:** Host informational sessions or webinars about your project. This can help attract interested participants and also provide an opportunity for them to ask questions and better understand the project.
5. **Outreach to Schools and Universities:** Work with educational institutions to reach students who might be interested in participating. This could involve presentations, information sessions, or collaborations with relevant departments or student groups.
6. **Job Boards and Volunteer Websites:** Post the opportunity on job boards, volunteer websites, or research portals. There are several online platforms where you can post opportunities for people to participate in research studies.
7. **Existing Networks:** Utilize your own existing networks and encourage them to share the opportunity within their networks. A simple email to colleagues, friends, and family can have a significant snowball effect.