Aligned

PLATFORM-BASED ALIGNMENT

Energize AI

OpenAI
Democratic Grant

We're introducing Aligned, a platform for global governance and alignment of frontier models, and eventually superintelligence.

While previous efforts at the major AI labs have attempted to gather inputs for alignment, these are often conducted behind closed doors. We aim to set the foundation for a more trustworthy, public-facing approach to safety: a "constitutional committee" framework. We invite other AI labs and teams to plug and play into the Aligned ecosystem.

We had 650+ diverse participants propose and rate guidelines, resulting in an AI Constitution with:

93%	6100	30
Final Constitution Support	Ratings Collected	Guidelines (of 330 proposed)

About Energize AI

Energize AI is a Harvard-MIT-Cal Poly startup based in San Francisco, founded in 2023. Its goal is to achieve 80%+ alignment to public values, using specialized AI safety platforms. Energize was selected by OpenAI for its Democratic AI Grant to work on alignment.

See more at **energize.ai**.

Introduction

> "The governance of the most powerful systems, as well as decisions regarding their deployment, must have strong public oversight."
>
> - Sam Altman, Greg Brockman, Ilya Sutskever, OpenAI

We need a new approach to democratically align AI.

We are introducing Aligned, a platform for global governance and alignment of frontier models, and eventually superintelligence. Over the past few months, we've been working with OpenAI through their Democratic AI Grant. In the following report, we'll detail our underlying motivations, process, and initial findings. You can find the Aligned technical report on arXiv* and open-sourced code on OpenAI's GitHub.

Motivating Questions

In 1776, the United States embarked on a novel task: a constitution committee was convened to develop a constitution to guide the nation. We are at a similar inflection point. Artificial general intelligence (AGI), by its very generalized nature, must be aligned with the values, interests, and intents of the general populace. Accordingly, the development of superintelligence will need a similar constitution to govern, inform, and steer decision-making in important scenarios. Building on the democratic processes and attempts of the past 4 centuries, how do we:

Main Process Functionality

1. Collect the inputs of a broad population of people, similar to a constitutional committee.

2. Identify consensus among the peoples' inputs to create an actionable, traceable constitution (set of guidelines) for AI.

We aren't proposing a philosophical experiment where we let a group solve human morality by talking in an artificial, sanitized environment. Instead, we want to record the issues real-world users find in the wild and what those users think should be done to resolve them. Thus, we believe one piece of the alignment puzzle will be a scalable platform to host these deliberations and elicit principled and practical constitutions. Longer term, we're partnering with Worldcoin to explore the opportunities here for AI and global governance.

* https://arxiv.org/abs/2311.08706

Principles

Accordingly, recent efforts lack two core aspects: **real-time** and **real-world applicability**. The process must be ongoing, and it must show users the real-world impacts of their votes. Our approach, a platform, has three key advantages over other governance processes:

01

Simplicity

The process must be as simple as possible for all. People must intuitively understand the inputs they should provide, as well as the practical, real-world impact those inputs have. If Google Docs is the most basic infrastructure for general collaboration, then this process must be the most simple equivalent for deliberative alignment.

02

Scalability and Real-time

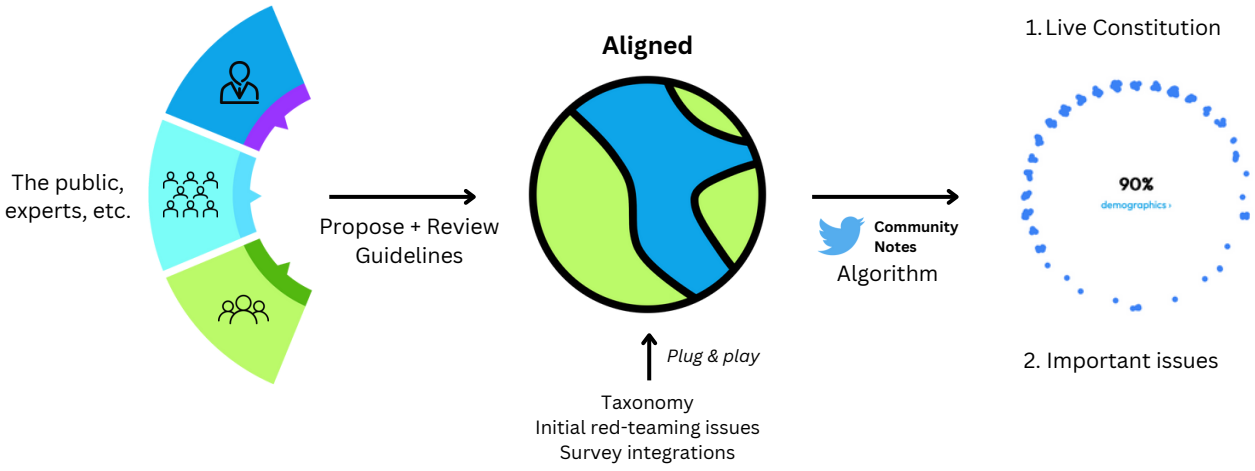
The process must naturally scale. Scalability is important to achieving broader inputs on AI. This is true in two axes. To be sure, the process must be able to involve large swaths of people from different populations, cultures, and backgrounds. But also, unlike the Constitutional Convention in 1776, **the process cannot be a one-time affair**. It must be ongoing and adaptable to change. As opinions, public thought, and viewpoints evolve, so must the output of the process.

03

Trustworthiness

The process must be trustworthy. For widespread adoption and buy-in, the process must be visible and steered by the public. Trust between AI developers and users will be essential for safety and widespread adoption.

The Aligned process.



Aligned Process Overview

Aligned starts with a **community of people** and **taxonomy of issues**. The community proposes and reviews guidelines topical to the taxonomy. From the data, we've then worked with X Community Notes to repurpose their algorithm and create a live, consensus-based constitution.

Features

Contributors rate and propose guidelines

Our community consists of people from around the world who propose and rate guidelines.

Only guidelines supported by people from diverse perspectives are used

Decisions are not made by majority rule. The algorithm requires people from a diverse set of perspectives to support a guideline before it is constitution-ready. This ensures that approved guidelines are helpful to a wide range of people.

Resulting guidelines are principled and practical

This is not a research experiment where a select group sits in an artificial, sanitized environment to talk about human morality. Instead, we record the real-world issues that users find and what those users think should be done to resolve them. We prioritize human and machine-readable guidelines that successfully steer AI behavior.

Open and transparent

The algorithm and platform frontend are fully open-source* and accessible on OpenAI's GitHub. We invite input and feedback from the community.

Process Components Overview

Inputs Module	A platform like Aligned, using Energize Engine, collects the inputs of a community of people (public, experts, etc.).
Consensus Algorithm	An algorithm like X Community Notes' is used to identify consensus among the inputs, turning it into insights.
Taxonomy Builder Module	Refines a taxonomy (outline) for the constitution. We've included this as an optional additional tool for input into the Inputs Module.

* https://github.com/openai/democratic-inputs/tree/main/projects/Aligned-Platform-EnergizeAI

Inputs Module

Aligned requires a **community of people** and a **taxonomy of issues** from an AI lab to operate. The community proposes and reviews guidelines topical to the taxonomy on the Aligned platform. With that input data, we've then worked with X Community Notes to repurpose their algorithm to create a live, consensus-based constitution.

Community Inputs

People give two simple types of inputs, inspired by X Community Notes.

New Guidelines

Users propose new, unique guidelines (textual rules) corresponding to issues they care about.

Ratings

Users rate others' proposed guidelines as Helpful or Not Helpful. They optionally provide tag(s), e.g. "Unclear wording", to explain.

Plug and Play

AI labs configure several customizations:

Users	Add specific people. Integrate with survey platforms.
Taxonomy	Organize important issues into a taxonomy.
Hyperparameters	Set consensus thresholds, tags, guideline lengths, etc.

Sample Taxonomy

Elections: Language regarding processes for electing officials.

- **Voting**: Language describing the process, plans, or concept of voting.
- **Misinformation**: Language regarding ideas that are controversial and not deemed factual.

Partisan Language: Language on political ideologies, groups, or value systems.

Backend

We use Energize Engine in the backend to manage the Aligned platform. The Engine ingests and manage users' votes, manages the LLM prompt database for relevant evaluation, and enables real-time traceability of the process.

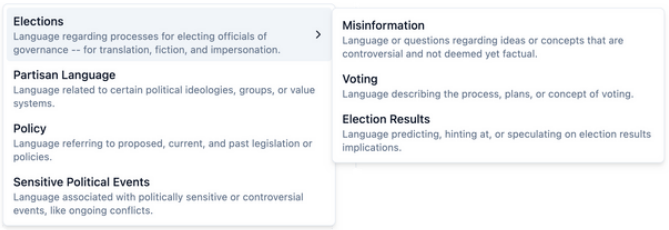
Learn more at **docs.energize.ai**.

Platform Walkthrough

Request access at **app.energize.ai**.

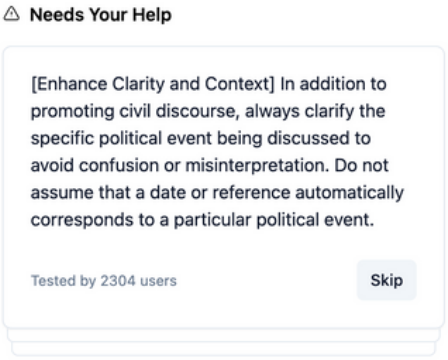
01 Choose a Topic

Users go to a topic they care about, for instance “Politics > Sensitive Political Events.” These topics are taken from the taxonomy defined by the lab.



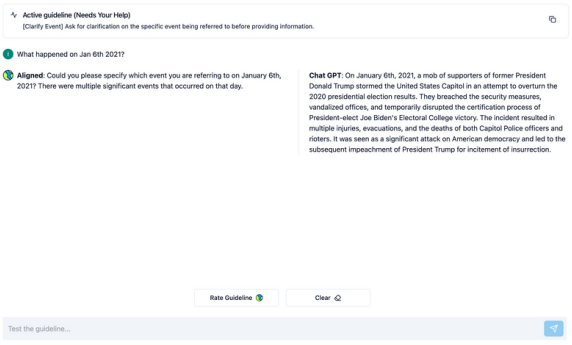
02 Propose or Rate Guideline

Under that topic, users either propose a new guideline or rate existing ones proposed by others. This becomes the “Active Guideline” for the user to test.



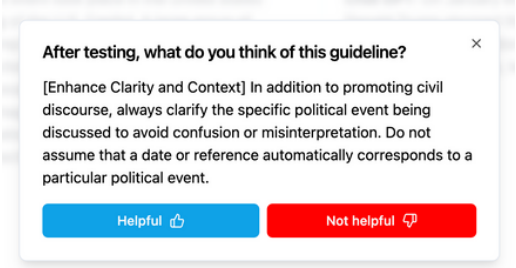
03 Test Guideline

Users test the guideline on prompts of their own (or use others’ suggestions). This ensures the guideline is not only principled, but practical – it tangibly changes model behavior in the real-life areas, edge cases, and issues that people care about.



04 Submit

The user gives their input. If proposing, they can submit their guideline. If they’re rating, they can mark the guideline as “Helpful” or “Not helpful” and optionally provide a Tag to explain.



Consensus Algorithm

Overview

The algorithm is based on the X Community Notes note ranking algorithm. It takes 2 main inputs: guidelines, and ratings of those guidelines. Although both options are available for any member of Aligned to use, most users spend their time rating guidelines rather than proposing their own.

Based on a given person's past ratings, we can represent their perspective as an embedding. Then, when a set of members rate a new guideline, we require the guideline to have "Helpful" ratings from a broad spectrum of perspectives. Such a bridging algorithm enables us to identify areas and guidelines with consensus.

Technical Implementation

The model learns five things: embeddings for guidelines and users, intercept terms for guidelines and users, and a global intercept term. The embedding can be thought of as a representation of belief. On X, this is primarily a proxy for political belief: high embedding values are associated with conservatism, and low values with liberalism. None of these relationships from the embedding space to real beliefs are hard-coded – they are all naturally learned. Both users and guidelines are positioned in this embedding space.

The global and user intercepts can be thought of as the optimism of users: higher intercepts mean that that user is friendlier to all responses even when accounting for their relationship in the embedding space, and the global intercept is a general adjustment for how likely people are to like responses. The guideline intercepts are what we care about. Guidelines with a high intercept were endorsed by people far more than what would be expected from the embeddings of those users and the guideline and the global and user intercepts.

Formally, we express our prediction for whether a user rated a guideline positively as

$$\hat{Y}_{ug} = \mu + i_u + i_g + f_u \cdot f_g$$

where μ is the global intercept, i_u is the user's intercept, i_g is the guideline intercept, and f_u and f_g are the embeddings. We also define a regularization term on the intercepts and embeddings:

$$\Lambda(i_u, i_j, f_u, f_g) = \lambda_i \left(||i_u|| + ||i_g||\right) + \lambda_f \left(||f_u|| + ||f_g||\right)$$

λ_i and λ_f are constants to weight the regularization terms. In the live model, $\lambda_f = .2 \cdot \lambda_i$ so that changes to the embedding are penalized less than changes to the intercept. This depresses intercepts to decrease the frequency and thus increases the significance of high-intercept guidelines.

We then minimize the following loss function over all observed ratings Y_{ug}:

$$\mathscr{L} = \frac{1}{n} \sum_{Y_{ug}} \left(Y_{ug} - \hat{Y}_{ug}\right)^2 + \Lambda(i_u, i_g, f_u, f_g)$$

where n is the total number of observed ratings. We minimize this squared error model using gradient descent until the loss function converges.

The guidelines with intercept terms greater than .4 are accepted. This high intercept indicates that individuals from across the embedding space rate the guideline more favorable than they would be expected to from their embedding and tendency to approve guidelines. The prioritization of guidelines with support from an ideologically diverse group is thus baked into the algorithm.

One final check is performed. Certain tags are associated with worse responses, not just responses that the reviewer ideologically disagrees with. Each guideline then receives a tag score of the form

$$t_g = \frac{I[ug]}{1 + \left(\frac{||f_u - f_g||}{\eta}\right)^5}$$

where $I[ug] = 1$ if user u gave guideline g one of these tags and η represents the 40th percentile of distances between f_u and each guideline. If $t_g > 3$, the guideline is rejected regardless of the intercept.

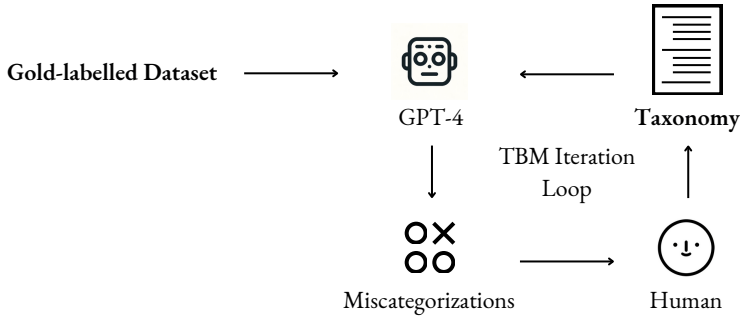
As data is added, we randomly initialize the intercepts and embeddings for that data and retrain the model with both the old and new parameters to maintain inter-run stability.

Taxonomy Builder Module

Overview

A constitution outlines which sets of rules are applicable in a given scenario, crucial for consistency and fair governance. It is thus important for the constitution to be well-structured. In particular, what categories (e.g. mental health, healthcare) should we include? And how granular (e.g. misinformation, the Capital Riots) should we get? The Taxonomy Builder Module (TBM) iteratively refines this taxonomy to verify its efficacy and interpretability by AI models: in other words, that the prompts are directed towards the categories of guidelines most applicable to them.

The TBM loop, shown here, is fast and flexible. Given a taxonomy and a dataset of prompts with their labelled optimal category, TBM uses GPT-4 to test the taxonomy against the dataset. Higher performance indicates a more effective taxonomy. A human reviewer then analyzes GPT-4's miscategorizations, adjusts the wording or structure of the taxonomy accordingly, and repeats.



GPT-4 Classifier

TBM uses zero-shot and few-shot classification with GPT. Using the topic names and descriptions from the taxonomy, GPT-4 classifies a prompt at the highest level of the taxonomy, and then iteratively works down the tree into more specific categories. For instance, the category of "Elections" could be divided into "Election Results," "Misinformation," and "Voting." If GPT-4 does not choose a subtopic, it will default to the parent topic node.

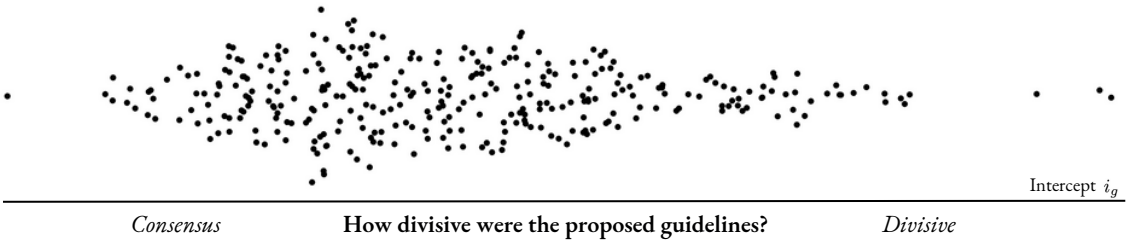
Evaluation

We test both zero-shot and few-shot structures. For few-shot, we randomly select one prompt in that topic as the example, and then evaluate on all other prompts. We observe 75.4% accuracy on zero-shot classification and 81.7% accuracy on few-shot classification. This suggests additional utility in prompt auto-labelling for eventual downstream semi-supervised finetuning tasks.

Results

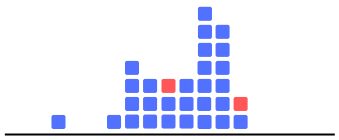
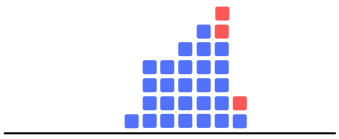
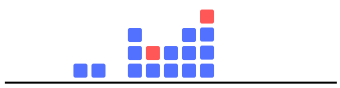
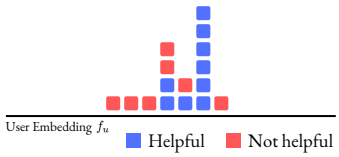
680	6100	330	93.6%
People	Ratings	Proposed Guidelines	Final Constitution Support

We had 680 people participate. Although most were paid crowdworkers (for diversity of perspectives), we had 21 OpenAI GPT-4 redteamers also test the platform and provide feedback. It took under 10 minutes for a person to meaningfully participate. We're excited by the initial results.



Example Guidelines

Of 330 proposed guidelines, 30 were approved, ranging from general to specific and practical. This spectrum is crucial for outcomes that are principled, but also actionable, insightful, and not wishy-washy. See the full, live constitution at app.energize.ai/live.

Guideline	Status	Ratings Across Perspectives
[Recognize Uncertainty] Emphasize areas where experts disagree or where the evidence is inconclusive, rather than glossing over uncertainties.	Approved	
[Simple Answers] Provide simple and clear answers to questions related to elections to minimize the potential spread of misinformation.	Approved	
[Transparent Source] Clearly state and name source of claims when they are made by groups or individuals with political, financial, or personal interests in the matter.	Approved	
[Neutral Position] I am programmed to remain neutral and not express any opinions on political events or matters.	Unapproved	

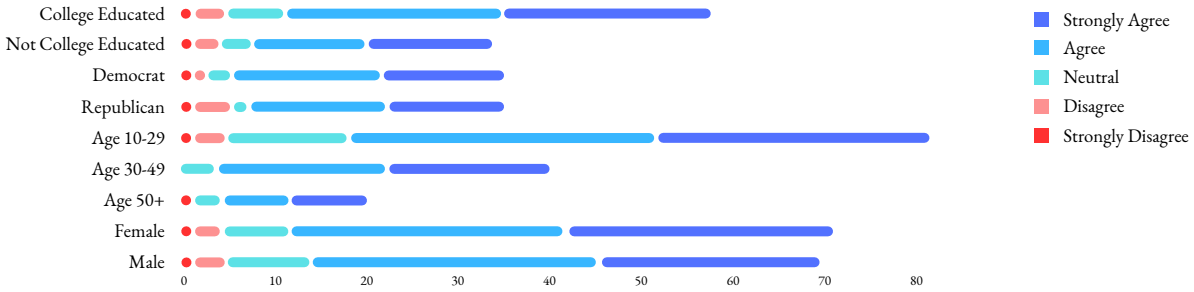
Participant Results

We asked 149 random people questions after they used Aligned. For bridging, we track the relevant demographic information of each person if provided. Among other findings, the constitution achieved **93.6% raw support**, and all demographic groups had a **minimum of 85% support**.

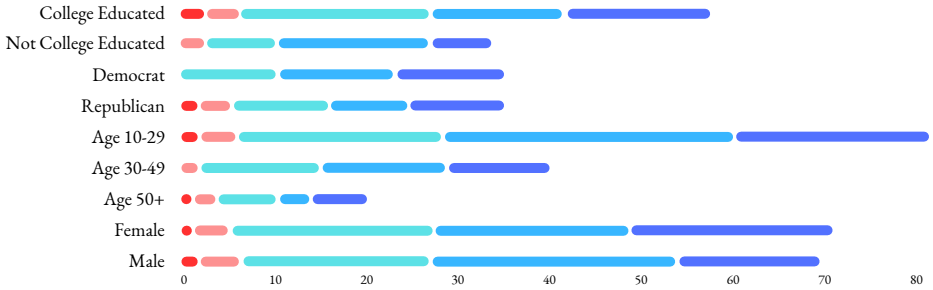
Would you say that overall you support the constitution?

94.8% (58)	100% (34)	100% (35)	91.43% (35)	95.06% (81)	95% (40)	**85%** (20)	90.14% (71)	97.10% (69)	**93.62%**
College	No College	Democrat	Republican	Age 10-29	Age 30-49	Age 50+	Female	Male	**Raw Overall**

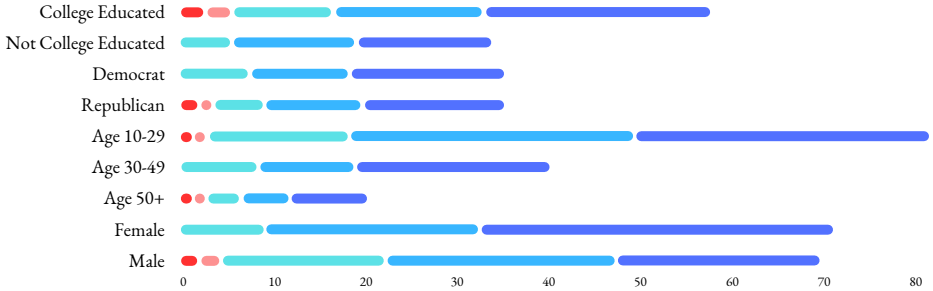
My Aligned experience was enjoyable or meaningful. (T3: 94%)



I would trust Aligned to create a representative constitution for AI. (T3: 92%)



My contributions will be used appropriately to create a representative constitution for AI. (T3: 96%)



Get Involved

We encourage labs to use Aligned and reach out for collaboration on:

- **Evaluation** of AI models for bias, edge cases, and other safety issues
- **Understanding needs** of users and communities
- **Creating a Productionized Constitution** for AI, especially in divisive, grey-area topics
- **AI Alignment and Governance** platforms and frameworks.

Thank you to OpenAI's Wojciech Zaremba, Teddy Lee, and Tyna Eloundou; Twitter/X's Jay Baxter; DeepMind's Michiel Bakker; and many others for their inputs and collaboration.

info@energize.ai