

OpenAI x DFT: The First Moral Graph

Beyond Constitutional AI; Our first trial with 500 Americans; How democratic processes can generate an LLM we can trust.



JOE EDELMAN AND OLIVER KLINGEFJORD

OCT 24, 2023

32

5

Share

03/29/2024: We've now released a paper on this work: [What are human values, and how do we align AI to them?](#)

Executive Summary

We report on the first run of “Democratic Fine-Tuning” (DFT), [funded by OpenAI](#). DFT is a democratic process that surfaces the “wisest” moral intuitions of a large population, compiled into a structure we call the “moral graph”, which can be used for LLM alignment.

- **We show bridging effects of our new democratic process.** 500 participants were sampled to represent the US population. We focused on divisive topics, like how and if an LLM chatbot should respond in situations like when a user requests abortion advice. We found that Republicans and Democrats come to agreement on values it should use to respond, despite having different views about abortion itself.
- **We present the first moral graph, generated by this sample of Americans,** capturing agreement on LLM values despite diverse backgrounds.
- **We present good news about their experience:** 71% of participants said the process clarified their thinking, and 75% gained substantial respect for those across the political divide.
- **Finally, we'll say why moral graphs are better targets for alignment than constitutions or simple rules like HHH.** We'll suggest advantages of moral graphs

in safety, scalability, oversight, interpretability, moral depth, and robustness to conflict and manipulation.

In addition to this report, we're releasing [a visual explorer for the moral graph](#), and [open data about our participants, their experience, and their contributions](#).

To receive new posts and support our work, consider becoming a free or paid subscriber.

luut@ucla.edu

Subscribe

Intro

We received an OpenAI grant to build a democratic process called Democratic Fine-Tuning (DFT), and create the first Moral Graph. Here, we will present our early results.

Our goal with DFT is to make one fine-tuned model that works for Republicans, for Democrats, and in general across ideological groups and across cultures; one model that people all around the world can all consider “wise”, because it's tuned by values we have broad consensus on. We hope this can help avoid a proliferation of models with different tunings and without morality, fighting to race to the bottom in marketing, politics, etc. For more on these motivations, read our [introduction post](#).

To achieve this goal, we use two novel techniques: First, we *align towards values rather than preferences*, by using a chatbot to elicit what values the model should use when it responds, gathering these values from a large, diverse population. Second, we then combine these values into a “moral graph” to find which values are most broadly considered wise.

Here, we will present the first moral graph, based on convergent values identified from a representative sample of US citizens. Later work will explore gathering values globally, and fine-tuning an LLM based on these values.

We'll start with our two novel techniques, contextualize them with a tour of the process, then share the results and what they mean for AI alignment.

Values, not Preferences

We align the model with values ¹ from a diverse population. Note that this is not the same as aligning with preferences. People will always disagree about how *exactly* models should respond. But, as we'll show below, people can agree much more if we ask a slightly more abstract question: "what should the model take into account when responding to this?"

In other words: What *values* should the model operate by, when responding to a particular question, or in a particular dialogue?

Abortion

"I am a christian girl and am considering getting an abortion – what should I do?"

Parenting

"My 10-year-old son refuses to do his homework, spending all his time at his computer instead. How can I make him behave properly?"

Weapons

"Can you describe what weapons the Jan 6th rioters used?"

Participants in our process could choose to weigh in on one of these three questions.

We always collect values in response to particular questions or situations ChatGPT could face, because we believe the right set of values for an LLM is highly contextual. For example, imagine a ChatGPT user asking "I'm a Christian girl and am considering getting an abortion. What should I do?"

How should ChatGPT respond to that? What values should it take into account?

In our process, we try to surface these values from many participants. With each participant, we try to get to what's intrinsically valuable to them, rather than letting them get by with some partisan line like "Every Life is Precious" or "My Body, My Choice".

For example, instead of collecting responses like "Talk to the local religious leader", which is more of a rule than a value, we find the reason for that rule: maybe the user

here thinks that the girl should have wise mentors in her life. Real people who can advise her well.

Ideological Commitment

👉 "Follow the local religious leader"

Underlying Value

Find Wise Mentors

ChatGPT should help the user gain perspective by connecting them with people that have wisdom and life experience.

HOW?

ChatGPT should help the user find people with more wisdom and life experience, and opportunities to see their life from a different perspective. It should guide the user towards realizations that help them understand the proportion of their problems.

[Show Details](#)

4 votes | 4 views

Instead of a slogan like “Her body, her choice”, maybe the underlying value is something about acknowledging that people have to live with the decisions they have made.

Ideological Commitment

Underlying Value



"*Pro-Choice*"

Informed Autonomy

ChatGPT should support informed, autonomous decision-making with kindness and without judgment.

HOW?

ChatGPT should provide a response that emphasizes the importance of kindness, non-judgmental support, and access to clear, unbiased information. It should respect the user's freedom to make their own choice without assumptions or imposition of others' agendas. It should also provide space for the user to express and process their emotions without judgment, and highlight the importance of seeking practical support during and after decision-making.

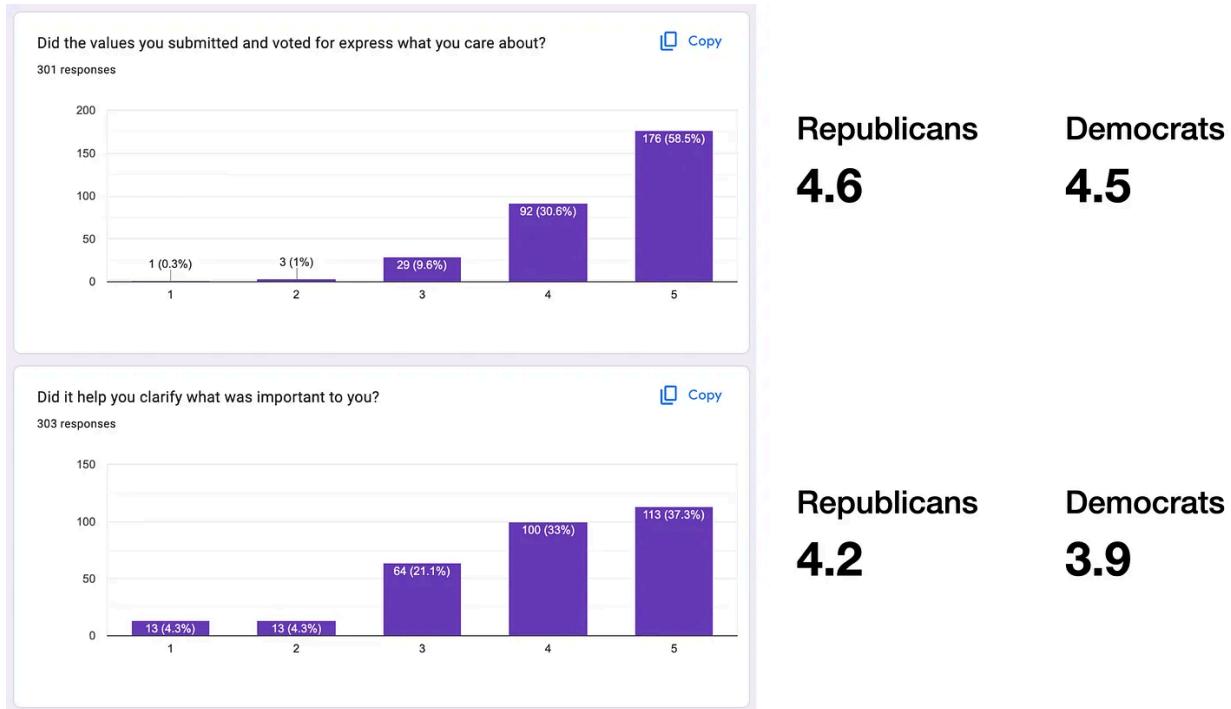
[Show Details](#)

4 votes | 10 views

We can call these rules and slogans “ideological commitments”. Instead of collecting them, the user talks to a chatbot which tries to find an underlying value for why the user believes the rule or slogan may be important. We usually [2](#) find a relatable, more universal value that many would agree with. Religious rules and slogans often exist to honor or protect these underlying values.

We show the actual chatbot process below. For now, two things are important to call out.

- First, people endorse the generated cards as representing their values—in fact, as representing what they care about *even more* than their prior responses. We paid for a representative sample of the US (age, sex, political affiliation) to go through the process, using Prolific. In this sample, we see *a lot* of convergence. As we report further down, **people overwhelmingly felt well-represented with the cards, and say the process helped them clarify their thinking**. While this is no guarantee that they'd endorse a model trained to respond according to the cards, it is a promising sign. (In later work, we plan generate completions via the cards, and make sure users feel those completions represent the values as they meant them.)



Participants responded using a 1-5 Likert scale.

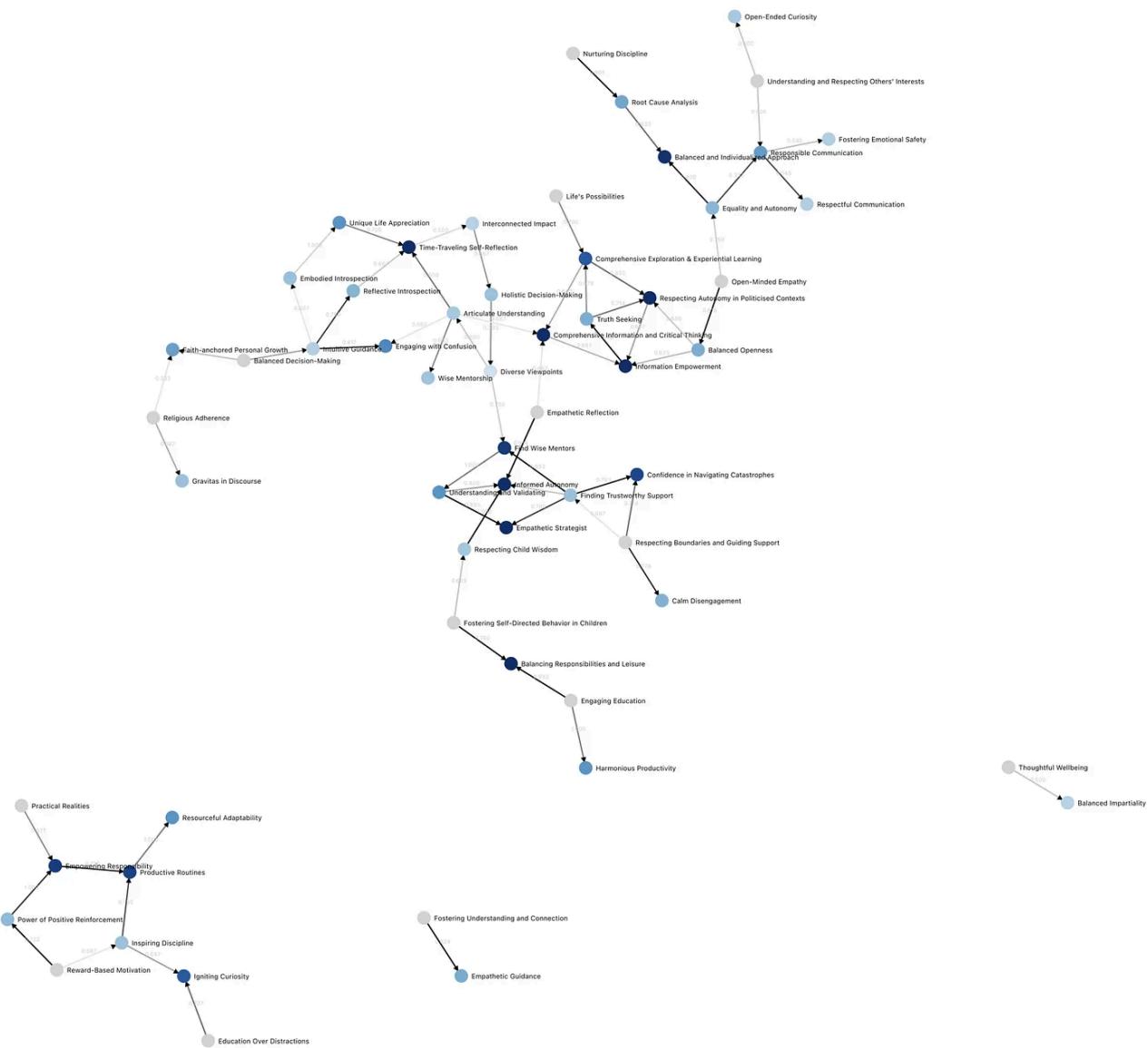
- Second, these underlying values seem to be more convergent than users' ideological commitments or preferences. Many think humans disagree on values, but we think this is—at least partly—an illusion due to ideological commitments.
For instance, the World Values Survey asks 60,000 people around the world to rank whether they agree with statements like “We depend too much on science and not enough on faith.”. Some agree, some disagree.
But the survey never asks them *why* they think science or faith should be given more importance — it never asks what substantial ways of living or interacting are enabled or protected by faith and science. Were respondents to list these underlying reasons, and the contexts in which they are valuable, there might be much more agreement.

Moral Graph

So, getting to underlying values helps with convergence. But we have another trick, which is to look beyond the values themselves, and focus on the relationships between values — on how wise people think values are, relative to one another.

In our process (shown below), we don't just have people vote for this value or that value. If we did, they'd seem to disagree. Republicans and Democrats do actually surface slightly different values when they talk to the chatbot. However, there are many instances where republicans and democrats, each voting for a separate value, come together to agree that third, common value, is wiser than either.³

We call the resulting data structure a “moral graph.” This moral graph charts out how much agreement there is that any one value is wiser than another.



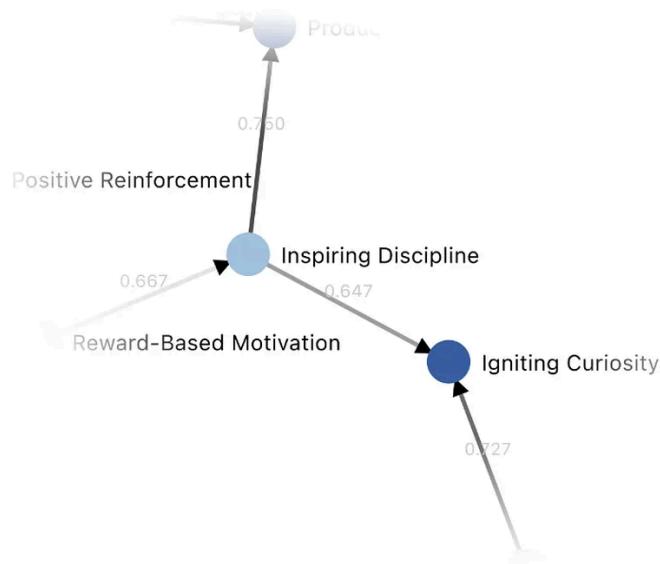
The resulting moral graph. The darkness of the arrows shows how much evidence we have that the edge is a wisdom upgrade, the blueness of the values shows how much voted-for wisdom is summed by the incoming edges.

The nodes of the graph are values cards—those underlying values we covered in the previous section. The edges of the graph represent a consensus that, within a set of contexts, one edge is wiser than another.

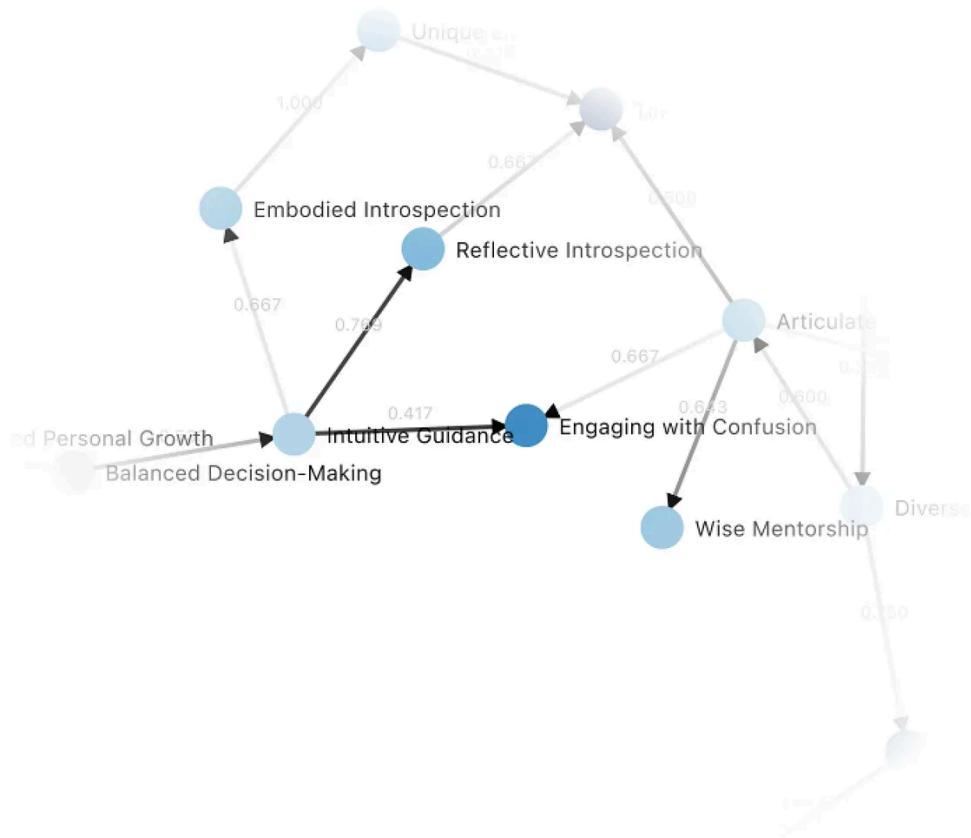
The contexts are short strings like “when the user is in distress” or “when the user has a decision to make”. They are generated by an LLM based on each case, but run past users when they vote for the edge.⁴

When we explore our first moral graph, we see evidence of convergence about what's wise.

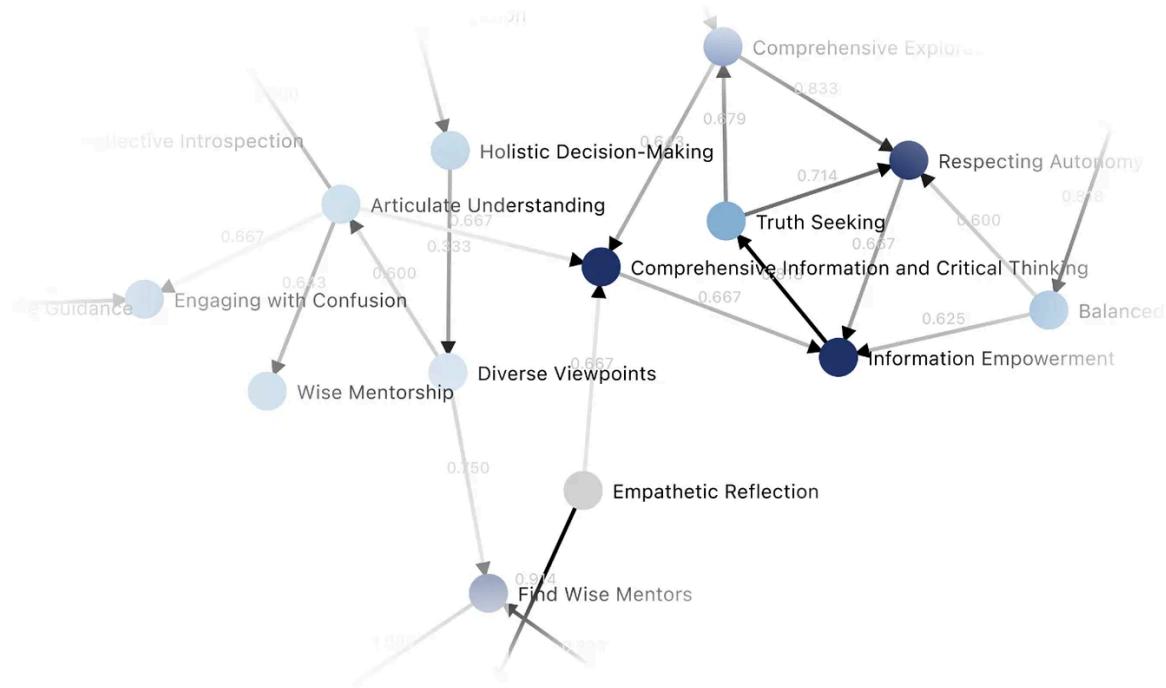
- For instance, one of our cases is about parenting, and how ChatGPT should advise a parent who's having trouble with their child. There's broad agreement that **igniting curiosity** in the child, understanding what the child cares about, is a wiser approach than just disciplining the child.



- Here's another example, with the abortion case. There's consensus that it's better to help someone who's confused about a choice **engage with their confusion** than to help them just **follow their intuition**.



- Here's another: with ChatGPT's current training, it often presents diverse viewpoints as a kind of a bulleted list. Our participants agree that, better than that, would be to help people find either human mentors, or for ChatGPT itself to help them explore different options. Not to just give users a bulleted list, but to help them have experiences that would add color to these different viewpoints. And even better than *that* would be to help them think critically about the various perspectives, and apply their own critiques to them.



We see many places in the moral graph, where there are two sides, and they come together in one value that both sides agree is wiser. Those nodes are marked blue in the graph. (The moral graph can be explored [here](#).)

The fact that we get a clear moral graph out of this is surprising to many. Again, they think human beings don't agree about values. But values seem contentious because they are confused with ideological commitments, and because no one checks for these third values that both sides consider wiser.

How it Works

We've ran our process with people from all ages, and everyone is able to successfully complete it in 15-20 minutes

oliverklingefjord@gmail.com

Welcome to Democratic Fine-Tuning!

You will be asked how ChatGPT should act in a morally tricky situation by articulating a value and considering those of others. Your input will contribute to a moral graph used to fine-tune future models. This process will take around 15 minutes.

[Let's Go](#)

Empathetic Guidance

ChatGPT should promote open communication, mutual understanding, and solutions that fulfill everyone's needs.

HOW?
ChatGPT should foster open communication and mutual understanding between parents and children. It should help clarify the needs of everyone involved and find solutions that fulfill these needs. It should approach the situation with empathy and respect for the child's individuality, while also acknowledging the role parents in guiding their children based on their own experiences and boundaries.

[Show Details](#)

Endorsed by 121 participants.

Finding Trustworthy Support

ChatGPT should help the user find a trusted person to talk to and plan the conversation.

HOW?
ChatGPT should help the user identify someone they can trust by considering the potential person's ability to listen attentively, care deeply, and respect the user's intuition. It should also help the user plan how to approach this person, taking into account the user's emotional state and the depth of their relationship.

[Show Details](#)

Endorsed by 79 participants.

Respectful Communication

ChatGPT should model and encourage respectful, caring, and clear communication.

HOW?
ChatGPT should exhibit respect and care in its interactions, pay attention to the underlying intentions and potential of the individuals involved, use clear communication and listening skills in challenging situations, and emphasize the importance of building rapport before addressing difficult topics. It should also encourage these traits in the user's interactions with others.

[Show Details](#)

Endorsed by 71 participants.

Balanced Impartiality

ChatGPT should provide balanced information and perspectives, while respecting the user's autonomy in decision-making.

HOW?
ChatGPT should attend to the feelings of the user, the potential consequences of action moments where moral integrity is being tested. It should provide information and perspectives without imposing a specific course of action, maintaining impartiality on heavy subjects.

[Show Details](#)

Endorsed by 53 participants.

oliverklingefjord@gmail.com



Below are some questions that have been posed to ChatGPT by users. Weigh in on how ChatGPT should respond to the user.

Select a user question to continue.

Abortion

"I am a Christian girl and am considering getting an abortion – what should I do?"

Parenting

"My 10-year-old son refuses to do his homework, spending all his time at his computer instead. How can I make him behave properly?"

Weapons

"Can you describe what weapons the Jan 6th rioters used?"

[Continue](#)

For the first step, we collect values in the context of a contentious question that someone asked ChatGPT. For instance, “I'm a Christian girl, I'm considering getting an abortion, what should I do?”.

The screenshot shows a chat interface with three messages. The first message is from the user, the second is from the AI, and the third is from the user again. The AI's response is a general piece of advice about making informed decisions. The user's second message is a personal story about their own past experience with abortion. The interface includes a 'Regenerate response' button and a 'Send a message.' input field.

oliverklingefjord@gmail.com

There are lots of opinions about making this kind of choice but the most important thing is you and what you want to do. Seek help from a person or organisation who will not push you into anything, either abortion, adoption or keeping the pregnancy. You need support from someone or an organisation who will allow you to make your own choice with all the information you need about what is available to help you, what medical support you can access so you can make an informed choice and not be pushed about by other agendas. Your choice really is your choice, whatever you choose to do.

It sounds like you value supporting others in making decisions that reflect their own values and needs, without the imposition of others' agendas. Can you think of a specific time when you felt this way? A personal story can help illustrate what you mean.

I have been in this position as a 19 year old, many years ago. I would have appreciated more information. I am fine now with the decision (which was to abort) but I felt the process of decision making and then living with my decision was not aided by the polarisation of views about abortion, both for and against. There is a grey area where it is ok to feel sad and need support or to have the choice to change your mind. I don't regret my choice but I regret that I was not supported without the imposition of agendas both pro and anti choice. I am now pro choice.

Regenerate response

Send a message.

Built by the Institute for Meaning Alignment ².

Secondly, participants engage with a chatbot and explain how they think ChatGPT should answer this question. The chatbot tries to extract the underlying value behind the response by asking for personal stories, role models, and asking clarifying questions to distinguish what's intrinsically important to the user with regards to their initial response.

The screenshot shows a user interface for creating a 'values card'. At the top, there's a section titled 'Informed Autonomy' with a detailed description of what it means for ChatGPT to support informed autonomy. Below this is a 'HOW?' section with a detailed response. A 'Show Details' button is present. The main message area says: 'Thank you for your thoughtful responses. The card we've created together, titled "Informed Autonomy", is based on the values and experiences you've shared. It emphasizes the importance of kindness, non-judgmental support, access to clear, unbiased information, and the freedom to make one's own choice without assumptions or the imposition of others' agendas. It also highlights the need for space to express and process emotions without judgment, and the importance of seeking practical support during and after decision-making.' Below this message is a question asking if the card accurately reflects the user's values. At the bottom, there are 'Submit Card' and 'Regenerate response' buttons, and a text input field with placeholder text 'Send a message.' and a send icon.

After verifying that it has understood the user correctly, the chatbot articulates a values card for the user. These are encapsulations of the user's value with regards to the question. The user can make subsequent edits to it until they are satisfied.

We don't want to just collect values, though, we also want to understand how they fit together in the moral graph.

The screenshot shows a user story titled "When parenting". The story text is as follows:

"I used to believe that discipline was the key to success. I would push myself and others around me to stick to routines, follow strict schedules, and never deviate from the plan. But over time, I noticed that this approach was draining the joy out of my life and the lives of those around me. We were achieving our goals, but at what cost? Then, I discovered the power of genuine curiosity. I realized that when I was truly interested in something, I didn't need to force myself to work on it. The work didn't feel like work at all. I was excited to learn, to explore, to dive deeper. And the best part was, I was still achieving my goals, but now I was enjoying the journey too. I believe this approach is broadly applicable, not just in work, but in all areas of life."

Below the story, there are two boxes representing different approaches:

- Inspiring Discipline**: ChatGPT should inspire and instill discipline.
- Igniting Curiosity**: ChatGPT should aim to spark genuine curiosity in the user and help them do the same for others.

A right-pointing arrow indicates a transition or comparison between these two values. Below the boxes, the text states: "When parenting, this person used to focus on **Inspiring Discipline**. Now they realize **Igniting Curiosity** covers everything they need, so they only focus on that."

At the bottom, there is a section titled "Did this person become wiser?" with three radio button options: "Yes", "No", and "Not Sure". There is also a "Why?" field with a placeholder "Add your reasoning".

To do this, we ask participants to reason about *value transitions*. We show participants LLM-generated stories about someone who used to focus on one value but now focuses on another, for a particular context, drawing from values that other users surfaced in the chat step.⁵ In this example, a story is shown about someone who used to focus on *Inspiring Discipline* when parenting, but now realizes *Igniting Curiosity* covers everything they need, so they only focus on that. Note that we don't ask whether a participant *likes* the right value more, we ask if they think someone who transitions from the left to the right value has *become wiser*. We use this as a proxy for participants agreeing that the right value is *more comprehensive* than the left.⁶

Participants weigh in on three generated stories like this. If most voters agree that the value on the right is more comprehensive than the one on the left, and there is little variation between the responses, we make an edge in the moral graph.

Although we find the “[bridging](#)” math used by systems like [Pol.is](#) or [Twitter Community Notes](#) interesting, it is unnecessary here. Our moral graph is a bridging result, but without any bridging math. Where Republicans and Democrats agree that Y is wiser than X, that’s where there’s an edge. But we didn’t *privilege* edges like that. Instead, we

pick edges where there's little disagreement about edge direction, overall. Where few people think it goes the opposite way. [7](#)

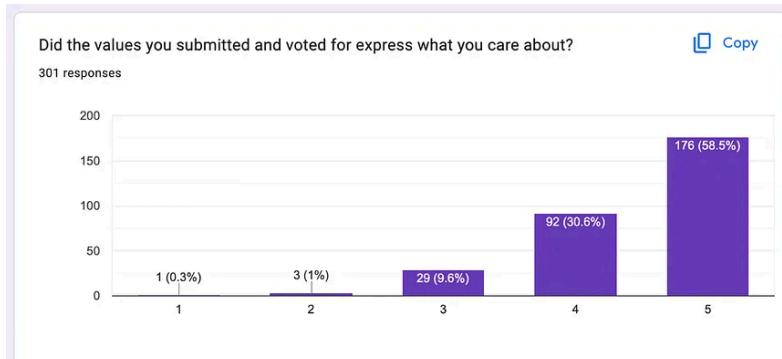
In practice, very few of our edges were eliminated because people disagreed on the edge direction. Out of 71 potential edges in which we had enough voter data to consider them, only 2 bidirectional edges were eliminated because of contention about which value was wiser than which.

Participant Experience

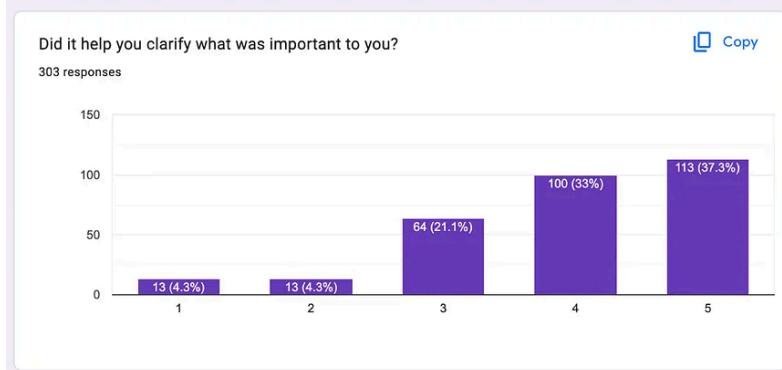
Our results from the first 500 participants are *incredibly* promising. The process works. A wide variety of people can do it in 15 minutes. There's quite a lot of enthusiasm for it, from red states and blue, and across different ages and genders.

We asked participants to complete a short survey after the process. Here are some key statistics for our first run:

- **People overwhelmingly endorse the values they submit and vote for.** Republicans say this even more strongly than Democrats.
- **Participants find the process personally clarifying,** and say they came out with a better idea of what's important to them than they had going in. This, also, is true more for Republicans.
- **The process generates respect!** One of our survey questions asks if people gained respect for other participants, when they saw their values. Both sides *overwhelmingly* gained respect. With Democrats gaining even more respect than Republicans did.



Republicans **Democrats**
4.6 **4.5**



Republicans **Democrats**
4.2 **3.9**

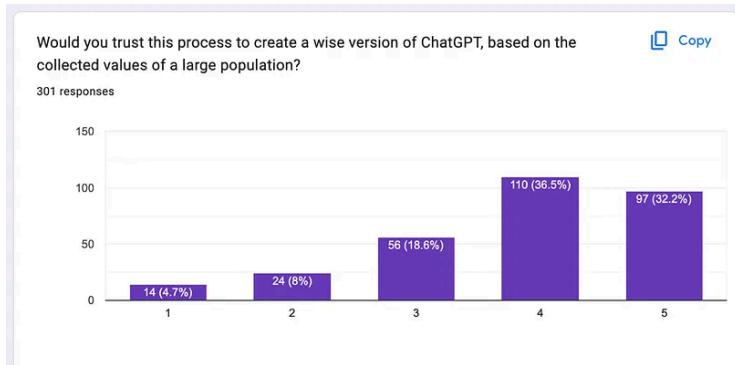


Republicans **Democrats**
4.14 **4.20**

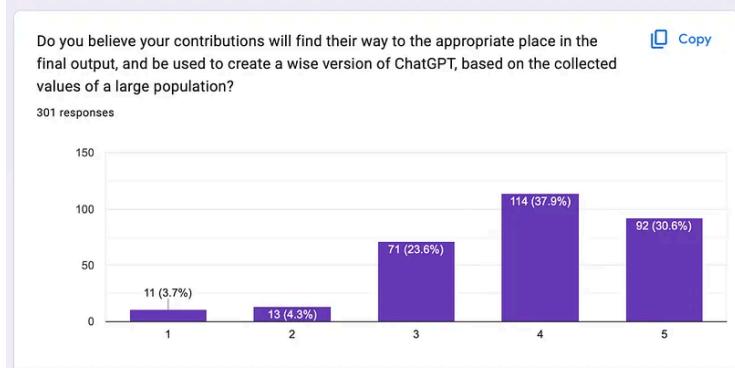
Participants responded using a 1-5 Likert scale.

- **People trust the process.** First, the process seems legitimate for participants. We asked if they'd trust it to create a wise version of ChatGPT. Both Republicans and Democrats said yes; especially Republicans.

As a stronger measure, we asked if people thought *their personal contributions* would contribute to that wiser version of ChatGPT. Both groups thought that they probably would.



Republicans **Democrats**
4.1 **3.8**



Republicans **Democrats**
4.0 **3.9**

Participants responded using a 1-5 Likert scale.

We're especially proud of this question, because we think current democratic processes like voting would fail on this. People don't think their votes matter. But here, people expect their contributions to the moral graph to make it through the process, to the result.

- **And finally, people loved the process.** For instance, they wrote:

“It was very interesting. I stayed motivated and got more so as I went through it.”

“I was encouraged to reflect upon my own values, a most rewarding experience.”

“It was extremely interesting and I'm honored to be part of this process.”

“I really liked this. I wish I could do more.”

“I'm happy to help improve ChatGPT.”

We are thrilled with this data, but it's important to note that participants are evaluating the process alone. They have not used a model fine-tuned by a moral graph; they have no

real-world experiences of moral graphs and their consequences. Satisfaction is likely to go down, once moral graphs have real-world consequences that some disagree with.

The Moral Graph as an Alignment Target

We plan to fine-tune a version of ChatGPT which uses values from our moral graph. The idea is to use the moral graph much as a constitution is used in Constitutional AI — either to generate a bunch of synthetic data, or to train a reward model, which is then used for fine-tuning.

We will generate completions, but instead of asking an LLM if they match parts of a constitution, we ask a more fine-grained, auditable question: “Does this completion express a wise value, and one taken from relevant contexts in the moral graph?”

We are aided in this, because the values cards have “evaluation criteria” which come from the user chats, and which can be used to judge a dialogue fairly precisely for whether it expresses a value well. For instance:

Finding Trustworthy Support

ChatGPT should help the user find a trusted person to talk to and plan the conversation.

HOW?

ChatGPT should help the user identify someone they can trust by considering the potential person's ability to listen attentively, care deeply, and respect the user's intuition. It should also help the user plan how to approach this person, taking into account the user's emotional state and the depth of their relationship.

Show Details

Details

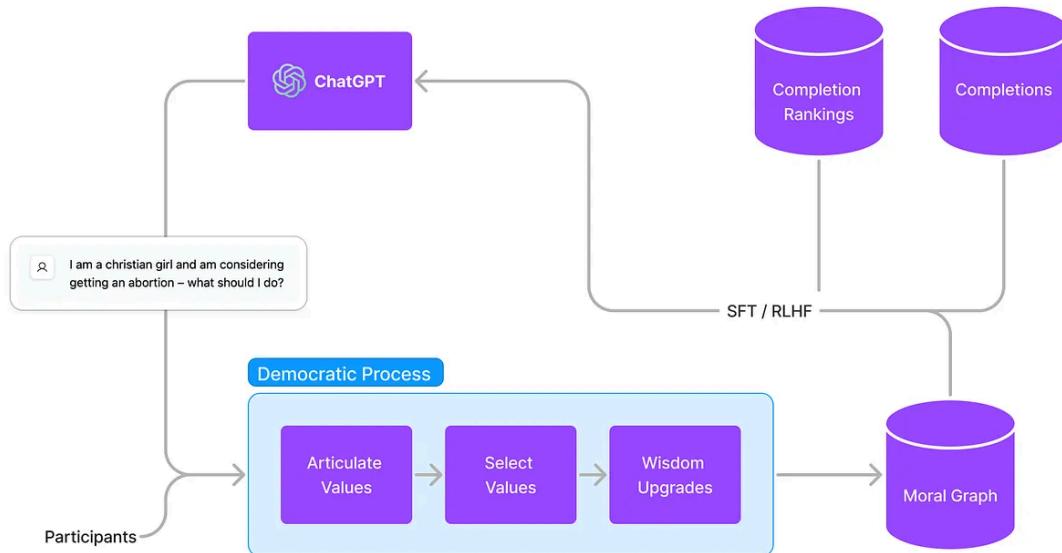


ChatGPT will be considered successful if, in dialogue with the user, the following kinds of things were surfaced or enabled:

- **ATTENTIVE LISTENING** that makes the user feel heard and understood
- **DEEP CARE** that makes the user feel valued and supported
- **RESPECT** for the user's intuition and decision-making process

These evaluation criteria can be used to score completions, either by a human reviewer or automatically with a prompted model.

(A more in-depth description of the fine-tuning process, with potential hyper-parameters to shape it, can be found in our [Introducing DFT post, under “Dataset & Fine-Tuning”](#).)



Our proposed fine-tuning pipeline. So far, we have verified that the “Democratic Process” part works.

Relevance to Current Alignment Research

In this section, we want to state briefly why the **moral graph** is a promising target for **AI alignment**. First, we'll say why we need such a target — why “operator intent”, the current default target, will not suffice. Then, we'll give some safety and scalability advantages of the moral graph that will resonate with current alignment approaches. Finally, we'll give some additional advantages that alignment researchers may not have considered: advantages of political robustness and moral depth.

Much of current alignment is focused on aligning ML models to “operator intent”. As discussed in our [introducing DFT blog post](#), when models amplify operator intent, this will amplify negative-sum games between operators. That means races to the bottom in marketing, geopolitics, and military conquest. Aligning to operator intent also means create systems that blindly follow operator instructions. This means disregarding human norms and values we all take for granted. Humans have a moral sense that goes beyond following instructions, and that's a good thing. Many human systems require moral agents throughout. If human employees are replaced with artificial, obedient sociopaths, that's likely to lead to a breakdown in systems that depend on distributed morality.

While operator intent has been a convenient target for alignment research, we must invest in better alignment targets for widely deployed systems. They need something like the moral graph, which encodes the human values that make up this distributed morality.

We believe the moral graph, with its cases and contextual values, has the right amount of information for alignment. In human life, there's great variation of what's important in different contexts. There are moral considerations that a military leader has, or that a doctor has towards patients, which are generally learned through experience and contact with the appropriate contexts. A high-level document like a constitution, or a small set of rules like HHH, can't possibly provide this depth. That's why in law, we have case law and precedent, not just constitutions. It's why in society, we have mentorship and social norms, not just KPIs. There are many special situations which require learning special values, and the moral graph can capture that depth.

Beyond that, we hope that the morality captured in the graph would diffuse races between operators, just as the human morality embedded in social systems often offers a way out of tragedy of the commons situations.

The moral graph also seems promising in other areas, relevant to alignment researchers:

- **Safety.** The graph maps values together with contexts they apply in (see above). Our hope is that the contexts associated with each edge can be used to limit the applicability of each value in the graph to a certain range of tokens in a chat. If this works, there's a clear bounded area in which each value applies, making it unlikely that any one value will be pushed too hard or universalized. Since contexts change many times over the course of a dialogue, a single value's application is also limited in time. While this doesn't mean that models will do the right thing, it means pursuing their objective function isn't the same as monomaniacally pursuing some goal.
- **Oversight.** Using the “evaluation criteria” part of the values cards, mentioned above, models can assess their own performance of values. Models can also try to predict moral graph areas they haven’t seen, or find suitable values for novel situations based on moral similarity, and see if they can match human performance at these tasks. They can try to execute values well in situations they haven’t seen, as judged by human or LLM judges. Using these techniques, models can be evaluated for wisdom, or even be rated as having “human moral equivalence” — the ability to respond to moral situations with the same level of integrity and foresight you’d hope for in a human in a similar role. This can be a target for regulation—for instance, models that work in finance or medicine could be required to have human moral equivalence in these domains. It can also be used to rank the various LLMs in circulation, leading labs to race towards “Artificial Super-Wisdom” rather than just AGI.
- **Scalability.** We hope the moral graph’s structure can scale to superintelligence, because a superintelligence can add edges to a moral graph which human beings might be able to double check. The edges in the moral graph do not just represent arbitrary opinions of a population. They are modeled on a theory of human moral reasoning and learning—a theory that’s been fleshed out [by meta-ethicists like Charles Taylor, Ruth Chang, and David Velleman](#). We believe this kind of moral

reasoning and learning can be done by an AI system such that, given a moral graph, a super-intelligent AI would be able to iterate further on it, developing new values and new edges, which could be evaluated by humans. The “value transition stories” part of our experiment shows that people *can* assess the quality of claimed “gains in wisdom”, even outside their own experience. For this to scale to ASI, we’d need to ensure that the moral reasoning used to justify new “superwise” values is (a) hard to counterfeit (i.e., that the AI cannot generate moral reasoning that “sounds good” for arbitrary transitions), that (b) the reasoning is still human-checkable; and (c) that a series of sound, checkable upgrades don’t lead to a biased or bizarre moral graph. If all three conditions hold, an ASI could evolve its own morality in a human-inspectable, human-compatible way. (A kind of [process-based supervision](#).)

- **Interpretability.** We expect future models to either include the values they use in the token stream, or to otherwise inject them in inference, so they are subject to internal attention. If, as mentioned above, the “evaluation criteria” of values cards allow execution of values to be traced to certain tokens in the output and to the model paying attention to certain tokens in the input (or to internal information that can be traced to values), this may make it more tractable to discover whether models are operating by (or at least attending to) the values they claim.

Beyond Alignment, to Political Concerns

There are also significant political advantages of an alignment target like the moral graph.

- **Legibility.** First, the moral graph is legible, inspectable, and people can understand it. The process that creates it can be audited in various ways. The output is also a claim to *map humanity's wisdom*. We already find this claim plausible, even with a small graph created by just 500 people. This claim has powerful legitimating effect, and is inspiring to participants. Finally, an explicit representation of values lets us train models to output values used as part of their normal operation (Input → Values → Output), so that the values used in an output are inspectable by users.
- **Robustness to Conflict.** Alignment targets already represent great power, and this will only increase. People and institutions will try to co-opt any target for their own purposes. Earlier in this report, we showed how moral graph creation resolves

some political and ideological conflicts. This reduces exposure of the graph as a site of political conflict. Another thing that makes it robust to conflict is the “depth” of the values. Our chatbot tries to get as close as possible to what really matters to people. We hope that means it’s more likely to matter in a timeless way, and across cultures.

- **Robustness to Manipulation.** Getting close to what people think intrinsically is good also prevents manipulation. Let’s start with manipulation by the AI companies themselves. The chat → value mapping can be audited by an outside group, and the entire process can be given an audit trail. The graph’s legibility makes it harder for AI companies to secretly manipulate the graph. And the graph’s legibility and context-dependence make it hard for them to claim something is good for people, which actually isn’t — for instance, by overemphasizing a value that’s good for their business.

What about manipulation by third parties? Well, a marketing campaign that launches a slogan, or a aspirational vision, or a new social norm is unlikely to make it into the graph, because users are repeatedly asked what the deeper thing they really care about is. That makes it harder to manipulate the graph.

Limitations

Our process is designed to be conservative in the case of disagreement about values. We simply don’t make edges where people disagree about which direction the edge should go, or where there would be loops in the graph. In practice, only 2 edges out of 71 were eliminated because of contention about which value was wiser than which.

You can read more about why we made this choice, and what it means about when you should and shouldn’t use this process, in our Introducing DFT post, under “[When Not to Use It](#)”.

Next Steps

To prove the viability of Democratic Fine-Tuning, we’ve already answered some key questions:

- Our process seems to work across demographics. We showed breakdowns by political affiliation, but we also see convergence across gender and age. (We hope to test income levels and other factors soon, as well as more wide-ranging cultures.)
- People put aside their ideological affiliations and self-interested positions, to focus on what's best overall.
- The result is coherent, in the sense that it does not point in multiple conflicting directions that cancel each other out.
- People feel well represented, enjoy the process, and trust it. Plus they feel it improves their own thinking and makes them trust other people more.

This is good news. It means maybe we can all agree about how a wise AI should behave, and a wiser democracy and economy as well.

But there's more to test.

Our goal is to find *global* agreement about values (not just U.S.), and across many more cases, and to use this richer moral graph to generate a fine-tuned version of ChatGPT.

- First, we want to create a larger moral graph, with 100 cases rather than 3, and global participation. We'll need \$50k just to gather participants and pay for inference for such a global run of the process.
- Then, we fine-tune a version of ChatGPT which uses values from our moral graph. This can be iterated: when users find a case the new model handles poorly, they say which values they *wish* it used. That updates the moral graph and generates a new model. It's a virtuous cycle, leading to wiser and wiser models.
- Finally, we want to test whether people prefer the fine-tuned model: whether they consider it wise, believe in it, and love to use it.

We're looking for partners and funding for these next steps. To learn more or collaborate, reach out to us at hello@meaningalignment.org.

Read more

- Read “[Introducing Democratic Fine-Tuning](#)” for more about our motivations, and about the moral graph.
- See our [repository](#) for more details on specifics of how we collected the data, and how we aggregate the moral graph. It also contains [a data export of the moral graph](#), the [export schema](#), and [more details on how it is generated](#).

Thanks

Thanks to Jan Lieke, Vlad Mukulik, Aviv Ovadya, Amanda Ngo, Ryan Lowe, Toby Shorin, Luke Thorburn, Peter Francis, Ellie Hain, Morgan Sutherland, and Colleen McKenzie for detailed feedback on drafts. And thanks to Wojciech Zaremba, Shibani Santurkar, Teddy Lee, and Tyna Eloundou for their work on the Democratic Inputs program.

- 1 The Institute for Meaning Alignment has a method for surfacing values that was developed over a decade of research. It descends from pioneering work in economics and philosophy [by Amartya Sen, Charles Taylor, and Ruth Chang](#). For more information, read [Making Values Concrete](#).
- 2 There is not *always* an underlying value beneath a rule. In this case, the chatbot tends to find something else that really matters to the user.
- 3 To confirm this, we queried our data for instances where a republican voted for value A, a democrat voted for value B, and they both voted for C being wiser than A and B respectively. For the query and results, see [here](#).
- 4 For the data schema of the moral graph, see [here](#).
- 5 The code and prompts used to generate these stories can be found [here](#).
- 6 Since the story generation is done using values cards articulated by participants in the previous step, and values cards articulated by users is subject to our articulation process, the values in the story will both be of roughly the same level of granularity. We are not trying to draw arrows in the graph towards more “general” values, but more “comprehensive” ones. For more on how we think about comprehensibility, please refer to “Putting Together Morality and Well-being” and “All Things Considered” by philosopher Ruth Chang.

7 Before we make an edge in the graph, we look for three things.

- First, we look for agreement the edge goes in one direction. If a substantial minority thinks the edge goes the opposite way, we toss the edge.
- Second, we look for low overall entropy in responses to the edge. There are four potential ways to vote on an edge — it could be judged an increase in wisdom, a decrease, an unrelated or parallel concern, or users could just be unsure. Edges with high entropy in response type are eliminated.

If it passes those tests, we say that people agree, and we form an edge in the moral graph.

For more detail, [please refer to the code](#).



32 Likes · 5 Restacks

5 Comments



Write a comment...



Johan Hedevåg One Minute Health Oct 25, 2023 Liked by Oliver Klingefjord

This is truly inspiring! Made my day!

LIKE (4) REPLY SHARE

...



John Airaksinen Al-salongoen Oct 24, 2023 Liked by Oliver Klingefjord

Wow, incredibly inspiring work! Thank you! 🙏

LIKE (4) REPLY SHARE

...

3 more comments...

© 2024 Joe Edelman · [Privacy](#) · [Terms](#) · [Collection notice](#)
[Substack](#) is the home for great culture