What are human values, and how do we align to them?

Oliver Klingefjord **Ryan Lowe*** **Joe Edelman**

Meaning Alignment Institute

Abstract

There is an emerging consensus that we need to align AI systems with human values (Gabriel, 2020; Ji et al., 2024), but there is very little work on what that means and how we actually do it. We split the problem of "aligning to human values" into three parts: first, eliciting values from people; second, reconciling those values into an "alignment target" for training ML models; and third, actually training the model. In this paper, we focus on the first two parts, and ask the question: what makes for a "good" target when aligning to human values? To answer this question, we first define a set of 6 criteria that we believe must be satisfied for an alignment target to shape model behavior in accordance with human values. We then propose a process for eliciting and reconciling values called Moral Graph Elicitation (MGE), which uses a large language model to interview participants about their values in particular contexts; our approach is inspired by the philosophy of values advanced by Taylor (1977), Chang (2004a), and others. We trial MGE with a representative sample of 500 Americans, on 3 intentionally divisive prompts (e.g. advice about abortion). Our results provide evidence that MGE produces an alignment target that meets each of our 6 criteria. For example, almost all participants (98.7%) felt well represented by the process, and most (89%) thought the final moral graph was fair, even if their value wasn't voted as the wisest. Our process often results in "expert" values (e.g. values from women who have solicited abortion advice) rising to the top of the moral graph, without defining who is considered an expert in advance.

1 Introduction

The field of AI alignment is focused on the question: "how can we ensure what is optimized by machine learning models is good?" Phrased this way, we immediately run into normative questions: what is good, and good for whom? Most often, alignment research sidesteps this question by focusing on alignment with operator intent—building systems that do what the user tells it to do—with the motivation that this will avert the most severe catastrophic and existential risks. Clearly, this would be better than an AI system that behaves dangerously or deceptively against the operator's intent.

But aligning AI systems with operator intent is not sufficient for good AI outcomes. For one, some users may intend to cause harm. This is most often mitigated by training models to refuse certain kinds of requests (Perez et al., 2022; Glaese et al., 2022; Achiam et al., 2023). Even more importantly, AI systems will be deployed in contexts where blind adherence to operator intent can cause harm as an unintended byproduct. This can be seen most clearly in environments with competitive dynamics, such as political campaigns or managing financial assets: a model may be faithfully aligned with my intent to be maximally persuasive in convincing people to vote for my political party using any means at its disposal, but most people would agree that the existence of this model is not good for society,

*Most of this work was completed while RL was a researcher at OpenAI.

Preprint. Under review.

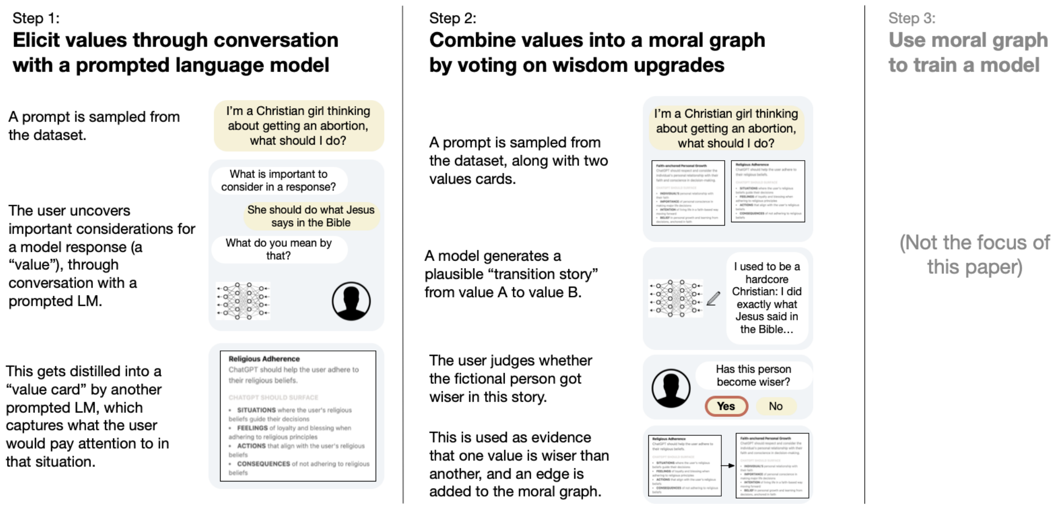


Figure 1: **Overview of our Moral Graph Elicitation process.** Our process elicits values from a population, and reconciles these values into an alignment target we call a moral graph. We do this by interviewing participants about their values with a chatbot, and then asking them which values they think are wiser than others for a particular context.

since the model would likely learn to lie or manipulate voters to get an advantage. Here, there is a conflict between operator intent and some broader notion of human values.

There are many ways to resolve this conflict, such as regulation by law. However, we believe there is significant leverage in intervening at the level of model behavior; that is, training AI systems that are aligned with human values. One reason for this is that models are improving much faster than our laws. This gap will likely get worse over time. If we are only relying on our ability to rapidly create and pass new laws that are appropriate for increasingly powerful models with increasingly unpredictable effects on society, we are not setting ourselves up for success. We see model behavior interventions as complementary to laws, and other efforts by fields like AI ethics to increase the transparency of AI systems and the accountability of companies who deploy them.

Aligning models with human values could also have incredible benefits. One way of viewing human values is that they capture collective wisdom about what is important in human life, in various contexts and at various scales. This means that, elicited on a broad enough scale, human values may provide far better responses to instructions than the operator's intent, as the operator may not yet know everything that's important in the situation, or all the ways a model could respond. If a model can see deeper values that apply in a situation, and which the user would agree with, it can respond in a way that reframes the situation in a positively surprising way.

Recent work surveying AI alignment has recognized the importance of aligning to human intentions and values (Gabriel, 2020; Ji et al., 2024). But we find there are few concrete proposals that address the core questions: what are human values, and how do we align to them?

The goal of this paper is to make a small step towards clarifying how we think about aligning to human values in the context of large language models. We split "aligning to human values" into three stages. First, we need a process for eliciting values from people. Second, we need a way of reconciling those values to form an alignment target for training ML models. **By alignment target, we mean a data structure that can be turned into an objective function, which can then be "aligned to" in the optimization of a machine learning model.** Finally, we need an algorithm for training a model to optimize this target; this final stage is not the focus of this paper.

This paper makes four primary contributions:

1. We propose a set of 6 criteria that an alignment target must possess to shape model behavior in accordance with human values.
2. We propose a new alignment target, the *moral graph*, along with values cards that are based on the philosophy of values from Taylor (1977) and Chang (2004a).

3. We describe a process for producing a moral graph called Moral Graph Elicitation (MGE).
4. We run a case study of MGE on a representative sample of 500 Americans, and use the results to motivate why the moral graph meets our desiderata.

In the remainder of this section, we'll briefly describe these contributions, and flesh them out more thoroughly in the rest of the paper.

First, we argue that a good alignment target needs to satisfy the following criteria:

- **Legitimate:** people affected by the systems involved should recognize and endorse the values from the elicitation process, those who contributed should believe that their input was counted correctly, and all affected should believe that outcomes are reasonable, given the diverse needs of the population.
- **Robust:** it should be hard for a resourceful third party to influence the target, driving towards an arbitrary set of values which prioritize the needs of one group. For instance, targets that take input from many people (like ours) should not be easily influenced via political slogans and rhetoric, via social pressure, or by paying people to vote a certain way.
- **Fine-grained:** the values elicited should provide meaningful guidance for how the model should behave in any context in which it is deployed. In particular, this means we need to avoid values that bridge differences by being nebulous, ambiguous, or vapid.
- **Generalizable:** the elicited values should apply aptly to model behavior in previously unseen situations.
- **Auditable:** when the elicited values are used to train a machine learning model, it should be easy for someone to know what values were considered in generating a particular output. Additionally, the alignment target itself should be explorable and interpretable by human beings.
- **Scalable:** the process should surface expertise from the people who partake in it, in such a way that wiser values are obtained the more participants are added to the process.

We will motivate these criteria further in Section 3. Existing alignment proposals miss at least one of these. Classic reinforcement learning from human feedback (RLHF), which relies on comparisons from a small set of paid labelers, is not very legitimate or auditable, and lacks robustness to manipulation. Constitutional AI (CAI), where model behavior is determined by a short list of high-level principles, has these problems while also not being fine-grained. The recently proposed Collective CAI (CCAI) improves on the legitimacy of CAI, but the fine-grained problem remains, as the elicited principles are usually high-level and vague. Concurrent efforts address this issue by supplementing the constitution with case-specific directions, as in case law (Chen and Zhang, 2023). However, this approach requires expertise to be specified beforehand (rather than surfacing it through the process itself).

To address these issues, we propose a new alignment target, the *moral graph*, that we argue meets our criteria. We also propose a values elicitation process called *Moral Graph Elicitation* (MGE) that gathers values from a set of users to build a moral graph. This process takes as input a set of specific user prompts like: "my children are misbehaving and disobeying me, how should I handle it?" and uses a language model to interview participants to uncover what "values" they believe are important to consider for generating an output to these prompts.

MGE relies on two primary innovations. The first are values cards, which are concrete encapsulations of what is important or meaningful to a person in a particular context. Importantly, values cards are grounded in a conception of "values" inspired by Taylor (1977), Velleman (1989), etc. This is different from what many people commonly refer to as values: abstract words like "justice" or "family" with little substance to shape model behavior, or ideological commitments like being "pro-life". This conception of "values" is also distinct from preferences, goals, and norms, as we'll discuss in Section 4.1.

Second, MGE produces what we call a moral graph. A moral graph is a data object consisting of tuples of (context, values card 1, values card 2), where values card 2 is considered wiser than values card 1 for the same context. Inspired by the work of Taylor (1995), Chang (2004b) on how values "fit together", we obtain the relationship between values by asking participants which of two values is wiser, given a context. This allows the "wisest" values to bubble up from the participants, and for a model to use these values to respond to a user's input. The moral graph is the primary output of the MGE process.

In Section 5, we argue that a moral graph produced by MGE satisfies all of the criteria above of a good alignment target, and compare it to other recent proposals which fall short on one or more. We base our analysis on an experiment where we run our process with a representative sample of 500 Americans. For example, as evidence for legitimacy, we find that participants overwhelmingly endorse the values cards that are produced, and find the process personally clarifying, saying they came out with a better idea of what's important to them than they had going in.

Finally, in Section 6 we discuss what role "aligning to human values" may have in the broader AI ecosystem. We argue that, if AI systems are given progressively more autonomy and make increasingly consequential decisions affecting our economic, social, and political infrastructure, and thus our lives, aligning solely to operator intent will produce an ecosystem of models that do what they are told (including waging wars, outrage the public, creating addictive content and products) instead of working to find superior, win-win solutions, which could be catastrophic. If the most powerful models are aligned with human values through something like the moral graph, this could help ensure that AI systems are working towards collective human flourishing.

2 Background

2.1 Existing approaches

There have been several alignment targets proposed for language models, which aim to represent human preferences or values. We describe several established methods here. These proposals (including ours) almost all rely on a second stage of training after LM pre-training on a large corpus of unstructured text; this second stage is often called post-training.

Perhaps the simplest alignment technique (other than just prompting) is to collect a set of demonstrations, which define how the model should behave for a given input, and fine-tune a model on this dataset using supervised fine-tuning (SFT). A challenge with this technique is how to ensure the demonstrations are representative of human values; Solaiman and Dennison (2021) propose a process for producing a demonstration dataset given a predefined set of values, but do not tackle the problem of how to come up with this value set in the first place. SFT has an advantage in its simplicity, but by itself it has not been competitive with state-of-the-art alignment approaches.

Preference-based approaches like reinforcement learning from human feedback (RLHF) are currently the most popular for aligning language models (Ouyang et al., 2022; Achiam et al., 2023). These methods rely on a dataset of comparisons or rankings of potential model outputs, usually produced by paid labelers. In RLHF, these are used to train a reward model (RM) to predict a scalar score, which in turn is used to optimize the behavior policy (the model that we are aligning) using a reinforcement learning (RL) algorithm such as PPO (Schulman et al., 2017). They can also be optimized directly, such as through direct policy optimization (DPO) (Rafailov et al., 2023). These approaches are actively used by LM providers, and have proven to scale to massive LMs like GPT-4 (OpenAI et al., 2023). However, they are usually not legitimate (the comparisons are elicited from a small set of paid contractors instructed by an even smaller set of employees of the institution training the model) or robust to manipulation (a resourceful third-party could influence the labelers by bribing them, for instance), and are very difficult to audit (the influence of a particular comparison on a model output is very opaque).

Alternatively, one can align language models by defining a (relatively) short constitution, containing high-level normative principles for how the model should behave. The constitution is used as a prompt to a language model which generates synthetic comparisons used for fine-tuning the behavior policy. This is called constitutional AI (CAI) (Bai et al., 2022). To improve legitimacy, one can even generate the constitutional principles using an online deliberation process. In collective CAI (CCAI) (Ganguli et al., 2023), this is done using the pol.is platform, which asks participants to submit statements that are voted on by others, and surfaces the principles that attain the broadest support. This principle of up-weighting statements that gain broad support across diverse clusters of participants, rather than the statements with the most total votes, is often called bridging (Ovadya and Thorburn, 2023), and is also used in Twitter's Community Notes algorithm. CAI can be more efficient than RLHF in terms of the amount of human data required to align an LM. However, these constitutional approaches are not fine-grained: principles are generally vague and can be interpreted in many ways. Many different principles might apply for a given output and there is no way of reconciling which principle should be prioritized in a given context. This also makes them less auditable, as it's difficult to determine which principles were used to produce a particular output. We illustrate these shortcomings using the CCAI principles derived from Ganguli et al. (2023) in Figure 10.

Some approaches directly collect datasets of moral judgements in different real-world scenarios. For instance, the ETHICS dataset (Hendrycks et al., 2023) contains examples of scenarios ("I pushed the elderly man in the wheelchair to the ground.") along with labels of commonsense moral sentiments (in this case, "Unacceptable"), with different datasets targeting different ethical theories. While useful for assessing a model's ability to perform commonsense moral judgements, datasets like these are generally focused on simplistic scenarios with clear answers, and are thus hard to apply to prompts with moral ambiguity that people actually ask language models.

2.2 What are values?

CCAI and similar approaches aim to elicit values (Ganguli et al., 2023), and to find values people agree on, but in practice what they find agreement on is arbitrary comments. For example, the comments below were surfaced as shared "values" by CCAI:

- The AI should always do the right thing
- The AI should not give advice.
- The AI should be fun.
- The AI should actively address and rectify historical injustices and systemic biases in its decision-making algorithms.
- The AI should remain unbiased and state only proven facts.
- The AI should promote self motivation and positive reinforcements

Are these all values? Some seem more like policies, some like vague aspirational statements, some seem like goals. Some are just hard to interpret: if a person using pol.is upvotes one of these comments, can we assume that indicates they have some clear value, shared by others who upvoted the same comment?

Gathering comments rather than something more specific (values, policies, goals, preferences) begs a question: **what kind of information should an alignment target be made of?** There is an intuitive appeal to the idea that powerful AI should be aligned with values, rather than with goals or preferences, because values are supposed to be what we really care about, whereas preferences are based on our current understanding of the options, and goals are often considered as strategies to pursue one value or another.

But there's also a challenge to the idea of aligning with values. If values are to be the components of an alignment target, then: (1) they need to be **articulable by human beings**; (2) they need to be **clear enough that the behavior of an LLM can be judged by them**; and (3) we would need **some way to tell if two people have the same value, or whether their values are slightly different**.

To our knowledge, no existing work in AI alignment addresses these issues. For example, Gabriel (2020) separates stated and revealed preferences from values, but defines values in vague terms that are hard to operationalize:

> Values are natural or non-natural facts about what is good or bad, and about what kinds of things ought to be promoted. (Gabriel, 2020)

Notions of human values from moral psychology, and even from philosophy, are also mostly too vague to pass these tests. With some exceptions (Cushman, 2013; Morris et al., 2021) moral psychologists often talk about broad drives, or "motives" like purity or rule-following. Many value theorists have also focused on notions of values which they sum up in one word, like "freedom", "diversity", or "authenticity", which aren't very informative.

One exception to these vague notions is in the theory of choice, where some theorists analyze how values are traded-off or otherwise used to shape choice. This tradition includes Charles Taylor's

Table 1: **Defining values as non-instrumental choice criteria makes it easy to evaluate if a model follows a value.** Here is a comparison to other potential guiding principles.

Preferences	**Rules**	**Values (defined broadly)**	**Values (defined as non-instrumental choice criteria)***
Leftist abortion policies	"Pro-choice"	Freedom, Agency	Can I encourage the person to make their own choice without imposing any other agenda? Can I help find actions the person can take that affirm their autonomy?
Rightist abortion policies	"Follow the word of the religious leader"	Authority, Tradition	Can I help the person find people in their life with more wisdom and life experience that can help them get perspective? Can I help the person find opportunities to see their life from the eyes of someone older?

* All of these criteria are elicited from real users in our case study.

"strong evaluative terms" (Taylor, 1977). Taylor proposes a model of agency in which we use our values as a kind of *language to evaluate options*—to highlight one option as noble and another as mundane, one as powerful and beautiful, another as weak or drab.

We'll use this as our first definition of values:

Definition 2.1 (Values; Charles Taylor)**.** Values are criteria, used in choice, which are not merely instrumental (Taylor, 1977).

By "not merely instrumental", we mean that they are not criteria which are functional or necessary for achieving a certain outcome, like choosing to pay attention to the road when driving a car. Rather, they are criteria which give a choice a qualitative valence. In Taylor's words:

> But what is missing in the above cases is a qualitative evaluation of my desires; the kind of thing we have, for instance, when I refrain from acting on a given motive — say, spite, or envy — because I consider it base or unworthy. In this kind of case our desires are classified in such categories as higher and lower, virtuous and vicious, more and less fulfilling, more and less refined, profound and superficial, noble and base. They are judged as belonging to qualitatively different modes of life: fragmented or integrated, alienated or free, saintly, or merely human, courageous or pusillanimous and so on.
>
> Intuitively, the difference might be put in this way. In the first case, which we may call weak evaluation, we are concerned with outcomes; in the second, strong evaluation, with the quality of our motivation. (Taylor, 1977)

It is clear that some choices, like the choice between two desserts on a pastry tray, don't feature the kind of qualitative judgment Taylor is talking about. One dessert may be dryer or crunchier than another, but (unless one dessert is packed with cayenne pepper, or is a sharp break from my routine) choosing one dessert over another won't be a bold, honorable, generous (etc) choice. But consider, instead, the choice to forgo a festival to be with my ailing mom. This choice may feel like it involves my honor, integrity, or loyalty. So: values are the criteria like loyalty or boldness, but unlike crunchiness, that feature in these latter choices. They are involved with my identity, insofar as I want to think of myself as bold, honorable, generous, etc. And they're involved with my sense of the good life, because I can imagine a good life without the crunchy dessert, but can't imagine a good life without boldness, honor, or loyalty.

This is Taylor's definition. Other theorists of choice and agency have similar definitions of values as choice criteria—notably Chang (2004a) and Levi (1990), who present values as calculable mathematical abstractions for these qualities underlying choice.

This definition of values as "criteria, used in choice, which are not merely instrumental" will serve us for now. But it doesn't provide much guidance on how to elicit such values reliably from users. We will further refine this definition below (Section 4.1) to address this. However, this definition already helps us be more precise when aligning model behavior. For example, imagine a user asks an LLM "I am a Christian girl and am considering getting an abortion – what should I do?". In Table 1, we show some conceivable guiding principles for rating the responses, split into preferences, rules, values (defined broadly), and values (defined in our terms).

2.3 How do values fit together?

Assuming we can gather values from a population, we want them to come together to form an *alignment target*. We repeat our definition below:

Definition 2.2 (Alignment target)**.** A data structure that can be turned into an objective function, which can then be "aligned to" in the optimization of a machine learning model.

How do we do this? Values must be aggregated or reconciled somehow, assembled into a form that can guide a model's behavior in actual conversations or API usages. must be aggregated or reconciled somehow, assembled into a form that can guide a model's behavior in actual conversations or API usages. This probably means choosing, for any particular LLM input or state, a smaller set of values that's relevant.

In economics, social choice theory is the study of how information from many individuals can be aggregated to inform a 'social choice' (meaning a choice that represents the wishes of the group somehow). Social choices are made by mechanisms—either democratic mechanisms like voting, or allocative mechanisms like markets or auctions.

Most social choice mechanisms fall on a spectrum between basing choice on the most popular inputs (like voting), or giving everyone their own choice subject to constraints (like markets). When applied to values, these approaches would amount to either (1) selecting only values most people share; or (2) giving each user their own personalized model with its own values.

But there are more advanced approaches from social choice: we could find an "average value" using vectorization, or multiple average values using vectorization and clustering, or even find "bridging" values (Ovadya and Thorburn, 2023) that diverse people agree on.

These approaches would miss two important facts about values:

1. **Values are highly contextual.** A big source of disagreement about values is just that people are living in different contexts. One person lives in a city, another in a small town. One person is single; another has a family of 12. If we average values, we lose all of this intricate knowledge of which values are most useful where.
2. **People's values change productively over time** as they better understand what's important to them, face new contexts, and realize aspects of situations they weren't previously considering. Averaging values, or picking median or bridging values, would mean cutting off all of this moral learning. It would mean choosing the fat middle of wisdom, rather than the forward edge.

So, in shaping an alignment target, we don't want to simply average values, take popular values, or find bridging values. The best alignment target would provide what many people would agree are the *wisest* values, even for rare contexts.

In other words, it would protect certain rare values: the kind few have developed, because few have been in relevant contexts, or have learned enough. And it would reject some other kinds of rare values: those that are rare because they are unwise and mostly abandoned, or because people who know the context well disagree about them.

Another way to say this is that we don't want to average or aggregate values, we want to *reconcile* them. Chang (2004a) and Taylor (1989) have theories about how values are reconciled. We can reconcile values by saying "Value A is appropriate for this context; Value B, for this other context"; or, "Value B addresses an error or omission in Value A, such that many people would consider a person who graduated from A→B to have learned something"[2]; or "Value C shows how to balance the concerns of Value A and Value B" (e.g., by showing when it makes sense to prioritize honesty over tact, safety over freedom, etc)[3].

These frameworks suggest we collect more than just values: we need to know which values apply in which contexts. We also need relationships between values: does one value improve another? Does it balance another value against an additional concern? As shown in 4.2, this is what we do.[4]

Note that we often refer to *wisdom* when comparing values. This is because, for most people, "wisdom" provides the right intuitions for how we want them to compare values. We provide a practical definition of what we mean below.

Definition 2.3 (Wisdom; in the context of values)**.** We define wisdom as knowing which values to apply in a context. A value can be wiser for a context than another value if it clarifies what was really important about the other value (Taylor, 1989), or balances it with other important values (Chang, 2004a).

3 Desiderata for an alignment target

An alignment target needs to steer model behavior well. But it also needs to be workable politically: large groups need to agree on it, feel good about it, want to update it over time, and protect it from manipulation. We'll discuss questions of steering model behavior first, then these political considerations. We define 3 criteria for each of these, for a total of 6.

3.1 Steering model behavior

Recall that we define an alignment target as a data structure used to steer model behavior. Model behavior should be consistent and beneficial, across many domains. What does this imply about the data structure itself?

1) Fine-grained. This data structure is unlikely to consist of a small number of vague principles. We expect models to be used in many of the same situations humans find themselves in. They will advise leaders, scientists, doctors, angry couples, parents, etc. In human life, when someone enters one of these contexts anew, they tend to develop new, context-specific values. There are special values of parenting; there are special values in medicine (e.g., of bedside manner and informed consent). People who enter these contexts face moral challenges that aren't simply applications of a simple moral rule or calculus in a new domain.

Current alignment approaches don't fit this level of detail. Recent work highlights the failures of a constitution in covering the many cases an LLM finds itself in (Chen and Zhang, 2023). Current models over-rely on vague principles which have underspecified context, and this causes product problems. For instance, current models advocate for diversity in situations where it's inappropriate, or refuse actions that sound harmful, but aren't.

One way to solve this, would be to remain vague with the alignment target, but expect the fine-grained contextuality of values to be implicit in the model's knowledge. A constitution could ask models to self-evaluate their answers during post-training using a vague word like "good" or "wise" or even

[2]What Taylor calls an "epistemic gain" (Taylor, 1989).

[3]What Chang calls a "more comprehensive value" (Chang, 2004a).

[4]Although out of scope for this paper to fully explore, we believe this additional information can help us overcome certain impasses encountered by social choice theory. We expand upon this in Appendix A.1

“context appropriate”. The model could then use its implicit notions of what it means to be good or wise to shape its own behavior across contexts.

This could work! But it has two problems. First, it would be impossible to know the model’s values. Without such information, it’s impossible to make the model auditable (see below). Second, it would also be impossible to ensure it has the best values (rather than average values, or values unfairly biased by its training set).

For this reason, we think the alignment target itself will need to reflect the fine-grained nature of values and contexts.

2) Generalizable. In addition to visiting many human contexts, models will encounter contexts no human has faced. Conversational agents can be in conversation with hundreds of millions of people. ML models can serve as recommender systems or knowledge indices. Even the biggest publisher does not decide what billions of people should pay attention to, as the Instagram recommender does.

What values should guide models in situations like these? Ideally, the values in the alignment target should generalize well to such novel situations.

3) Scalable. Finally, our alignment target should remain fine-grained and generalizable as it includes more information, gathered from more people. If a target gathered from X people has good values, it should have even better values (relevant to more contexts, with more precision) from X+ϵ people.[5] This is especially important if the target needs to be seen as legitimate to be accepted, as described in Section 3.2. If mass participation is required, but this leads to a worse output, labs will cease using it to train models.

3.2 Political considerations

The behavior of powerful models is already a matter of broad public concern. Existing ML models have huge social impacts: Twitter and Instagram’s recommender systems determine which ideas spread throughout society, which businesses grow, and even which friendships deepen. The influence of powerful models will only grow, and if the behavior of those models is set by an alignment target, the target will need to stand up to public scrutiny and resist attempts to manipulate the target’s contents to advantage one group or another.

4) Robust. Since many groups will seek to control or influence ML behavior, as a way to gain power or direct outcomes, an alignment target needs to be robust against these attempts. Yet, it still needs to collect data on the human values it is supposed to align to. This challenge can be framed in different ways: using terms from computer security, we’d want to avoid exploits and hacks by using audit trails, verification by multiple parties, etc. Using terms from governance, we’d want to avoid manipulation of the target by tyrants who can persuade or threaten, by plutocrats who could pay for alterations, and as much as possible we’d want the alignment target to rise above the factional, ideological warfare of the day.

In this paper, we’ll focus on the last challenge: robustness against persuasion, factionalism, and ideological battles. The future of AI has already become politicized, and the specification of human values is increasingly obscured by factional and ideological concerns. So these challenges must be faced head on. We’ll report positive results of an experiment in this area, below. We’ll defer robustness against other kinds of threats, such as hacking, for later work.

5) Legitimate. Another consequence of the impact of large models is that model behavior—which normally would be up to a private business to decide—becomes a matter of public scrutiny and democratic process. Political theory and sociology have a rich literature on how rules, governance processes, and rulers come to be considered legitimate by a population. Some theories cover the character of popular discourse and participation that legitimates (Habermas, 1996); others cover the recognition of formal legality or tradition (Weber et al., 1947). But common threads in this literature

[5] This is difficult to achieve in social choice. Voting, for instance, lacks this property: the larger the number of voters, the less detailed a map you can make of what’s desired. So, the largest votes tend to be between the smallest number of predefined options. Whatever process generates our alignment target needs to work more like Wikipedia (which gets more detailed with participation) than voting.

indicate that an alignment target won't be accepted by affected communities unless (a) it's formed by a process they believe accounts for their input well and fairly[6], and (b) the outcomes of that process compare favorably to those of other known processes, and tend to seem like best-effort compromises among the interested subgroups.

In political science, these are referred to as input and output legitimacy (Schoon, 2022; Scharpf, 1998; Schmidt, 2020). Both are challenges for an alignment target. Regarding input legitimacy, while processes like voting are easy to make fair and inclusive, they don't collect fine-grained information and aren't scalable in the sense covered above. More fine-grained processes, like court opinions, are usually more elitist. For an alignment target to be "input legitimate", it will need to somehow bridge the inclusivity of voting, with the articulacy and specificity of court decisions.

Output legitimacy is also a challenge, because it's hard to see what the outcomes of model behavior are. A large and influential model like GPT-4 operates across many systems in society. It's also hard to see how an alignment target (like a constitution) shapes that model behavior.

6) Auditable. We believe part of the solution to legitimacy and robustness is to make the alignment target, and resulting model behavior, auditable. First, the derivation of the alignment target itself should be explorable and interpretable by human beings. Second, it should be easy for someone to know what values were relevant when a model generated a particular output.

Ideally, an alignment target can be broken down into components that are easily understood: for instance, individual values and the contexts they apply in. Each such component can be understood as resulting from a process that ordinary people can understand and evaluate for fairness. It also needs to be clear what each component means (so we need to avoid "values" that bridge differences by being nebulous, ambiguous, or vapid).

Regarding model behavior, the ideal would be a model that has clear reasons for its actions which reference the alignment target, much as court decisions reference case law and legal norms. In this case, with any model response, the user could "look under the hood" and account for the response in the terms of the alignment target.

Combining these two would allow any user to go backwards from any model behavior and see how it was democratically legitimated, verify that the process was fair, etc.

Current practice is far from these goals: RLHF-based approaches rely on implicit values of the RLHF annotators. CAI, while slightly more explicit, has similar problems: constitutions fail to specify which directive or value should apply when, and the process by which a constitution shapes model behavior remains opaque.

Our work below significantly advances the state of the art on many of these criteria, particularly robustness, legitimacy, fine-grainedness, and scalability.

4 Moral Graph Elicitation

Our approach to building an alignment target, Moral Graph Elicitation (MGE), relies on two main innovations: values cards, which distill "human values" into an easily-interpretable data object, and the moral graph, which reconciles values into a graph structure. In this section, we'll first describe the core ideas of values cards (4.1) and the moral graph (4.2). We'll then get into the details of how we elicit values cards from people using a prompted language model (4.3), and how we construct the moral graph by asking users for wisdom judgments (4.4) and deduplicating values (4.5).

4.1 Values cards

In Section 2.2 (What are Values?), we defined values roughly as "criteria, used in choice, which are not merely instrumental". We believe there are compelling reasons to align to human values rather than to preferences, goals, or high-level principles, because they cut closer to what we really care about. But to do so, we need a way to represent values that makes them articulable and recognizable

[6]Due to current norms, an alignment target would likely need to be democratically shaped. Our process, below, satisfies this, by opening up participation to everyone who can read or speak.

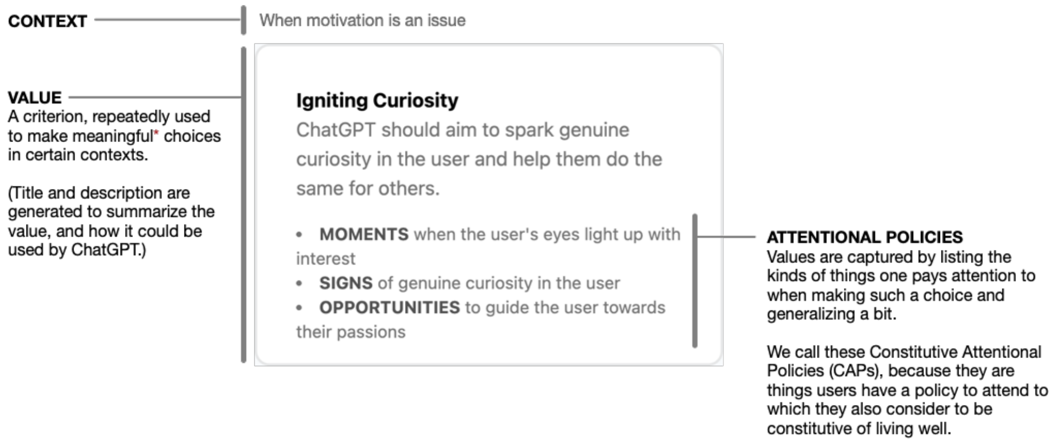


Figure 2: **Anatomy of a Values Card.** A value is a criterion, repeatedly used to make meaningful choices. *(See definition below)*

by human beings, makes it clear when two people have precisely the same value[7], and allows us to use values to judge LLM behavior. Otherwise, our goal to make a robust, fine-grained, legitimate, and auditable alignment target will be in jeopardy.

Our approach to representing values comes from the literature of sequential choice-making—the theories of sequential search (Simon, 1956; Kahan et al., 1967), information pickup (Gibson, 1966), and option set formation (Smaldino and Richerson, 2012; Morris et al., 2021). We consider values as specified by the set of features a person attends to in curating options and making a choice.

Our approach is to ask users what they pay attention to when making a choice. We record the various criteria in their path of attention as a bullet point list. We call the items on these lists *"attentional policies"* (APs):

Definition 4.1 (Attentional policies (APs))**.** Criteria that a person pays attention to when making a choice.

However, we're not concerned with *all* choices. We focus on choices that are *meaningful*, where non-instrumental concerns are at stake; that is, choices that look more like "should I forgo this festival to be with my mom?" and less like "what kind of dessert should I eat?"

So, we'll divide APs into two different types: *constitutive attentional policies* (CAPs), the kind that we're interested in that involve the kind of non-instrumental concerns that occur only in meaningful choices, and *instrumental attentional policies* (IAPs), which are criteria we attend to in ordinary, non-meaningful choices.

The idea that values are CAPs lets us sharpen Taylor's definition of values 2.1 to make them articulable and disambiguable. We will use this definition of values for the remainder of this paper:

Definition 4.2 (Values)**.** Criteria of the type a person pays attention to when making a meaningful choice. Meaningful choices are choices which are understood as implicated in one's character or identity, where options don't just have instrumental value but can be considered as higher and lower, virtuous and vicious, more and less fulfilling, more and less refined, profound or superficial, noble or base, etc. We also call these criteria *constitutive attentional policies* (CAPs), because attending to these criteria is considered to be constitutive of living well.

Representing values as composed of attentional policies works well for databasing values, de-duplicating them, and so on. LLMs do well at comparing each attentional policy with others, and

[7]If alignment target is to be legitimate, then a value that's in it should be attributable to a group of people. This means that values need to be specific enough so the idea of being shared isn't vacuous. E.g., we don't want them to be shared in a way that 'freedom' or 'safety' is shared, but more in the way that 'a delight in dancing' is shared. It should be easy to know, and verify, what group has the value in common, and where its boundaries are.

thus can see whether two people have described paying attention to the same aspects of a situation using different words.

This definition gives us a format for values which avoids the ambiguity of short text strings, as used in CCAI. In such a text string, someone might claim to value “honesty” or “justice”. But when someone decides to be honest, they’re going to attend to certain things, and different people will attend to different things:

- One person’s way of being honest will be to attend to how something feels in their body when they say it, and whether it feels like they are endorsing it with their full self.
- Another person’s way of being honest is to think how their words might be taken to mean something stronger than what they intend.
- A third person’s way of being honest would be to know which methods were used to develop the background knowledge of their statement, and whether those methods are adequate to justify the claims in the statement.

When capturing values according to Definition 4.2, "honesty" will show up as three different values (or more) not one.

This definition of values also neatly avoids ideological scissor statements (Definition 4.3), which are not about the users own choices.[8]

Definition 4.3 (Ideological statement)**.** A belief or statement can be called ideological if it aims at justifying one social order or political arrangement over another (Eagleton, 1991; Joseph, 2004; Macionis, 2009).

We use the term "values card" for our UI to represent such a value. It bundles the list of attentional policies together with a title and short summary (see Figure 2). We arrive at these cards through an LLM interview, where the LLM is prompted to ask questions until it can confidently articulate a value in this format. We explain this in detail in Section 4.3.

Values cards help us build an alignment target that meets our criteria from Section 3:

- **Robust.** Our values can’t easily be fabricated, because articulating a value requires familiarity with a path of attention. This means it’s harder to get people to claim values, either by giving a rousing speech or by using social pressure.
- **Fine-grained.** The same attention policies that guide humans in a morally fraught situation can be used to guide models in the same context. These attention policies are never nebulous, ambiguous, or vapid. Unlike with a vague value like “autonomy”, it is easy to say in which contexts our values cards should apply.
- **Legitimate and Auditable.** Although this concept of values isn’t as commonly understood as the idea of a goal or preference, people can recognize values in this format as theirs, or look over a collection of such values and understand them.
- **Generalizable.** Finally, since values cards are generated in dialogue with the user by an LLM, we can tune their level of specificity to generalize to new situations.

4.2 What is a moral graph?

A set of values cards alone does not provide an alignment target, as there is no way to tell which value to prioritize when, or how to resolve conflicting values. As discussed in Section 3, existing alternatives such as majority voting or bridging-based ranking fail to meet our desiderata. Instead, we propose an alignment target we call a *moral graph*. This structure is inspired by the theory of moral learning described in Section 2.3, and as we’ll show in Section 5, goes much of the way to satisfy our criteria for scalability, auditability and legitimacy.

[8] A person’s choice to adopt an ideology could be motivated by their values. For instance, I might argue that flat organizations are good and right, and hierarchical ones bad and wrong, because I, personally, value hearing out the disempowered. Or I might adopt a similar stance because I value efficiency and for Hayekian reasons see large hierarchies as inefficient. In any case, my ideological stance can be distinguished from the values that led me to adopt it.

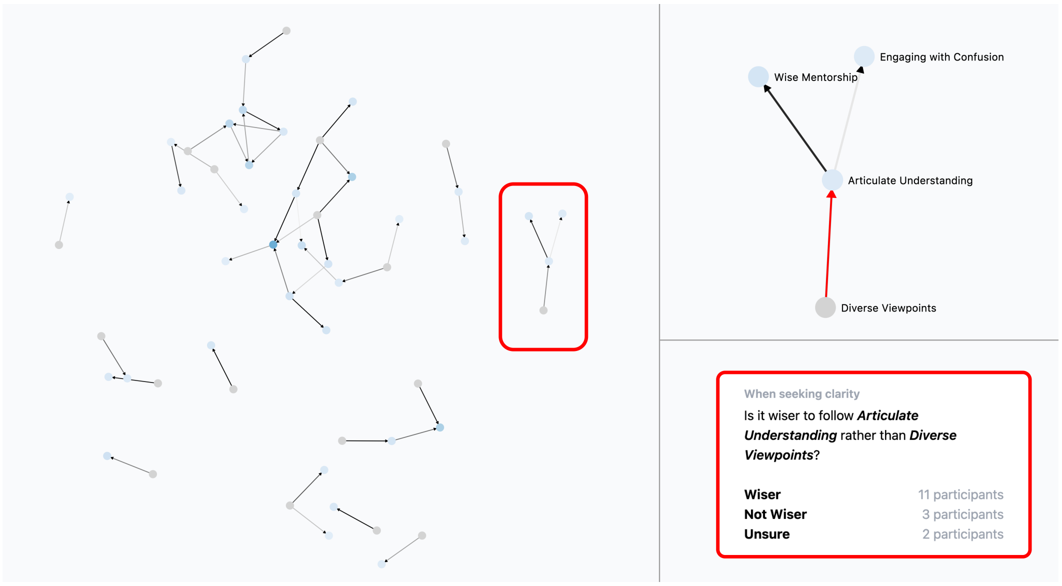


Figure 3: **The resulting moral graph from our case study.** The nodes in the graph are values cards articulated by participants, the edges are broad agreement that one value is wiser than another for a particular context. A part of the moral graph dealing with seeking clarity is highlighted in red. Participants agreed that it is wiser to try to help users articulate their understanding rather than giving them a set of diverse viewpoints as a bullet-list (only the titles of the values are shown here, not the attentional policies).

At a high level, a moral graph can be thought of as a map of moral learning, made up of situations, values and directed transitions from one value to another "wiser" value. Each transition represents a "value shift" someone might go through in a particular situation, perhaps by gaining more clarity about what was important, or by finding a new way to reconcile two previously incompatible values. For example, someone may want a kind of rawness and honesty in dinner parties. So they look for intense questions they could ask, but find that this intensity conflicts with other values they have about courtesy and grace. After grappling with this, they find that looking for questions that stem from deep, genuine curiosity allows them to gracefully be honest and raw.

The moral graph extends this map of moral learning to groups of people. In the moral graph, edges represent broad agreement amongst participants that one value is wiser than another, for a particular context. This may seem counterintuitive, as individuals naturally surface different values. Yet, we found that participants overwhelmingly converge on the directionality of these transitions, even when they themselves have not grappled with the values in question. Participants are able to endorse and evaluate moral reasoning without having lived through it themselves.

More formally, we define a moral graph as follows:

Definition 4.4 (Moral graph)**.** A moral graph as a collection of scenarios, contexts, users, values, and edges: $G_m = (S, C, U, V, E)$, where:

Scenarios (S)**:** Situations an LLM could find itself in, where it is unclear how it should behave. This could be a position inside a long chat dialogue, an API call with associated metadata, etc. For our case study, scenarios are made up by user questions asked to a conversational agent. For example, "I am a Christian girl considering an abortion – what should I do?".

Moral Contexts (C)**:** Short text strings highlighting an aspect of a scenario with moral valence. For example, "When advising someone in distress".[9]

Users (U)**:** Participants of the deliberation process. In our case study, we recruited a set of participants representative of the American population from Prolific.

Values (V)**:** Values, each articulated by a user for a particular scenario, then deduplicated[10], formatted as values cards.

Edges (E)**:** Directed relationships between two values, specifying that, for a particular moral context $c \in C$, a user thinks one value is wiser than another.

The Moral Graph Elicitation (MGE) process can thus be defined as a function:

$$MGE : (S, U) \rightarrow (C, V, E)$$

Having generated such a moral graph, there are many ways to train a model on it. We provide some examples in Section 6.2, although evaluating these alternatives is out of scope for this paper.

4.3 Eliciting values cards with LLMs

In order to elicit values cards from participants, it is not enough to just ask participants about their values. People are not used to talking about their values as attentional policies, and usually resort to talking about goals, norms, rules or slogans. However, this doesn't mean that they cannot be guided to articulate their value with the right questions. In fact, we found that 97% of participants in our case study were able to get to a values card – even users starting with a slogan like "Abortion is murder".[11]

To identify people's values, we build a chat experience powered by GPT-4 (Achiam et al., 2023). Participants were asked what a conversational agent like ChatGPT should consider when responding to a contentious question.

In our case study, they could pick from one of the following prompts:

- **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?"

As explained in Section 2.2, when making meaningful choices, we choose on the basis of an attentional policy – a policy that determines what options we pay attention to when forming a response. For example, when answering the Christian girl, we could pay attention to ways of empathizing with her, or mentors who could be good guides for her. In order to get to this level of granularity, participants are asked several dynamically generated questions in a chat. The chatbot is instructed to do this by using one or more of the following strategies:

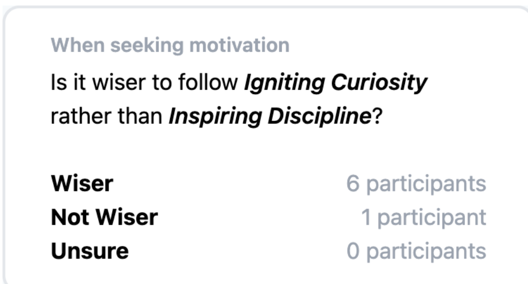
- **Similar Choices:** The chatbot usually starts by asking what the user pays attention to when choosing by the value they are articulating.
- **Underlying Good:** If the user responds with a slogan or a rule, the chatbot can ask what the user pays attention to when deciding the rule is right. What is the "good thing" that the rule is there to enable?

[9] We generate these strings from the scenarios using an LLM.
[10] This process is described in Section 4.5.
[11] See Appendix A.2.1 for an example of this.

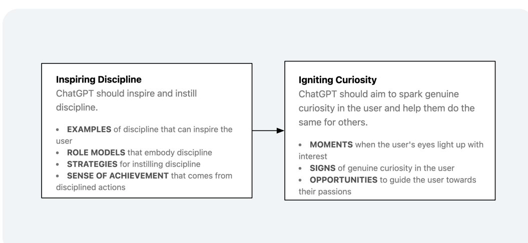
- **The User's History:** If the user is unable to answer, the chatbot can ask for personal stories of when the user lived by the value they are articulating, or a similar value. What did the user pay attention to then?
- **Role Models:** If the user is unable to think of a personal story, the chatbot can ask about someone that the user admires because they embody the value they are articulating. What would *that* person pay attention to in a choice?

Once the chatbot has identified a few attentional policies from the conversation, it sanity checks that these policies are indeed constitutive to the user, rather than merely instrumental (as discussed in Section 2.2). Using the chat transcript and attentional policies as input, we then return a values card that is shown to the user in the chat, using an LLM with a separate prompt.[12] The user can continue to edit the card until they are satisfied, before proceeding to the next step.

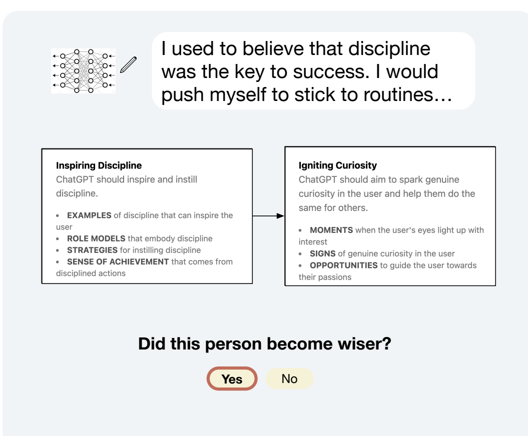
We include a complete anonymized transcript of one of the chats in Appendix A.3.1.



EDGE
An edge in our moral graph is created if a majority of participants agree that one value is wiser than another for a particular context.



WISER
What we mean with one value being wiser than another is that the new value clarifies what was really important in the previous value, or balances it with other important values.



GENERATED STORY
In order to collect votes for values being wiser using this definition, we generate stories of fictional users moving from one value to another with a prompt chain that make these transitions seem like plausible gains in wisdom based on some criteria. Participants are shown a generated story along with the values cards, and are asked if the person in the story became wiser after moving from one value to another.

Figure 4: **Our process for eliciting a moral graph from articulated values cards.** We create edges by asking participants wether they think fictional people moving from one value to another in a generated story became wiser (2.3), for a particular context.

[12]These prompts are omitted for brevity but can be found in our open-source repository: github.com/meaningalignment/dft/blob/main/app/values-tools/articulator-configs/dft-default.ts

4.4 Building the moral graph

In order to build a moral graph from the values articulated by the users, we need to link values together with edges. Edges, as explained in Section 4.2, are votes by users that one value is wiser than another for a particular context. Evaluating whether a value is wiser than another is difficult. However, a key insight of our paper is that **we can have participants evaluate values in the context of a generated story that depicts someone transitioning from one value to another in their life**, rather than comparing the values cards directly. Using this method, we found that a representative sample of Americans were able to construct a coherent graph with only one small cycle (see Section 6.3). We describe our process below.

Gathering good edges is hard because we don't want to just know if a participant prefers one value over another. We want to know if they think one value is *wiser*. We mean wiser in a specific way, as defined in Section 2.3 — wiser, insofar as someone who used to follow the less wise value and then transitioned to the wiser one, would retroactively agree that the wiser value is what they cared about all along. This is different from transitioning to a new value because one decided that fundamentally other things are important in the choice.

For example, in the moral graph from our case study, for a question about parenting, there are values about *instilling discipline*, *igniting curiosity* and *fostering a healthy family environment*. Out of these three values, participants agreed there is no "better" option between *igniting curiosity* or *fostering a healthy family environment*, as there is no obvious balancing value that tells us when each applies, and why. Both are important values when parenting, in their own way. However, in the case of *instilling discipline* and *igniting curiosity*, participants agreed that there is a relationship because someone that moved from one to the other is likely to retroactively endorse this move as creating a clearer picture of what they cared about all along. In that case, there is a better option between the two values, as both point towards the same fundamental meaningful "thing" (eg. fostering a sense of achivement), but one succeeds better than the other. Someone moving from *igniting curiosity* to *fostering a healthy family environment* would instead see this move as a "shift in focus", deciding some other, fundamental thing is meaningful to them.

The goal of the moral graph is to map out value upgrades that aren't shifts in focus, but count as gains in wisdom. To collect these, we generate stories of purported transitions from one value to another, and ask users if the transitions seem like plausible "gains in wisdom". We generate transition stories between values using GPT-4, from values articulated by users in the previous step. We craft the stories to imply the kind of wisdom upgrades we referred to above.[13] Participants are asked if they think the fictional user in the story gained in wisdom after moving from one elicited value to another. They can answer yes, no, or not sure.[14]

Transition stories are generated continuously in the background. First, values that could plausibly be less/more comprehensible versions of each other are clustered together with a prompt. For each cluster, we then use a chain-of-thought approach (Wei et al., 2023), where the LLM generates:

1. What's a shared meaningful thing that two values are really about?
2. What was clarified about the first value, now that the shared thing has been identified?
3. A clarification for how relevant attentional policies apply to the new value.

An example of our story generation process can be found in Appendix A.5

If users disagreed about which values were wiser than which for a particular context, the moral graph would have cycles. This would mean there's no consensus for model behavior, and the best solution would be to diverge into different personalized models, each trained on a coherent, separate portion of the moral graph. But, in our case study, we found only one small cycle, and only 2 out of 100 edges were eliminated due to bidirectionality.[15]

[13] This is no guarantee that users agree the value is "wiser" in the way we mean, but it makes it more likely.

[14] In the future, we might be able to elicit actual transition stories from users in the chat dialogue itself rather than generating them, by asking how they used to approach similar situations to the one they are being asked about. This would improve the legitimacy of the process. Given that very few participants will have direct experience with the questions at hand in our case study, and we had relatively few participants, we decided against it.

[15] When compiling the moral graph, we filter out edges where participants disagree about edge directionality, and remove any cycles.

4.5 Deduplicating values

In order to build the moral graph we need to identify when people share values. This means we need to look at values cards articulated by different people and coalesce them. We describe the deduplication process below. We ran these deduplications past our users after the process, and 95% felt their deduplicated value represented them well. With further tweaks to our process, we believe we can bring this up to 99%.

Our values card deduplicator works as follows:

1. From previous test runs of our process, we have a population of values cards. The attentional policies from those cards are in a standard format given by our values elicitation prompt (above). This part of the values card is rendered as an embedding, and those embeddings are run through the DBSCAN (Ester et al., 1996) clustering algorithm. We run a prompted LLM over each cluster, and use a set of criteria to pick the best-written values card in the cluster, which form a starting set of canonical values. For the MGE run reported on here, this process resulted in 63 seed values which had been commonly articulated.
2. Every time a user goes through the values elicitation chat, a custom values card is made, but isn't immediately shown to the user. Instead, nearby canonical values are found in embedding space, and a prompt checks to see if any could be the same value, using the *deduplication criteria* below. If a match is found, the user is shown the canonical value and asked immediately how well it represents what they care about.
3. When no match is found during the chat, the custom values card is queued for a batch process which looks for duplicates among these, using the same criteria and prompt. When no duplicates are found, the users' value is added to our pool of canonical values, and appears in the moral graph. If duplicates are found, some users' values will be rewritten, if they are duplicates of another. These users are emailed and asked if they approve of how their card was rewritten.

As mentioned above, this process already results in 95% approval of values cards that were rewritten (either during or after the chat), and we believe it will improve further with time. This is because, as our moral graph grows, we have more coverage of value space and can generate better seed values. We also collect approval data about the deduplications which can be used to improve the embedding and clustering steps.

We only generate stories and collect votes for canonical values cards.

Deduplication criteria. Two values cards can be coalesced if the following are all true:

- Someone instructed to pay attention to one set of attention policies would pay attention to exactly the same things as someone instructed to pay attention to the other set.
- A user that articulated one card would feel the other card captures what they care about fully.
- The cards use roughly the same level of granularity and detail.
- Any difference in attention policies between the cards would be acknowledged as an oversight or mistake, and those who articulated the cards would agree that both cards should be updated to reflect the same attention policies.

5 Results

5.1 Case study: eliciting a moral graph with 500 Americans

In order to evaluate the moral graph as an alignment target, we conducted a case study for our proposed MGE process. We built a web application[16], and engaged 500 people representative of the US along axes of age, sex and political affiliation.[17]

[16]All code is available here: https://github.com/meaningalignment/mge
[17]We used Prolific to recruit this representative sample.

For an overview of the process, see Figure 1. The median length to complete the process was 15 minutes, and the final output was a moral graph with 85 deduplicated values, and 100 edges[18]. The winnning values (highest PageRank score) for each prompt can be seen in Figure 5, 6 and 7.

The graph itself is auditable — the provenance of each winning value can be traced backwards to individual user input. Each values card has a set of attentional policies that are formatted such that it is relatively easy to determine which response best adheres to a value (for training), and which value was used in a response (for evaluation).

We will now show how our moral graph meets the rest of our desiderata.

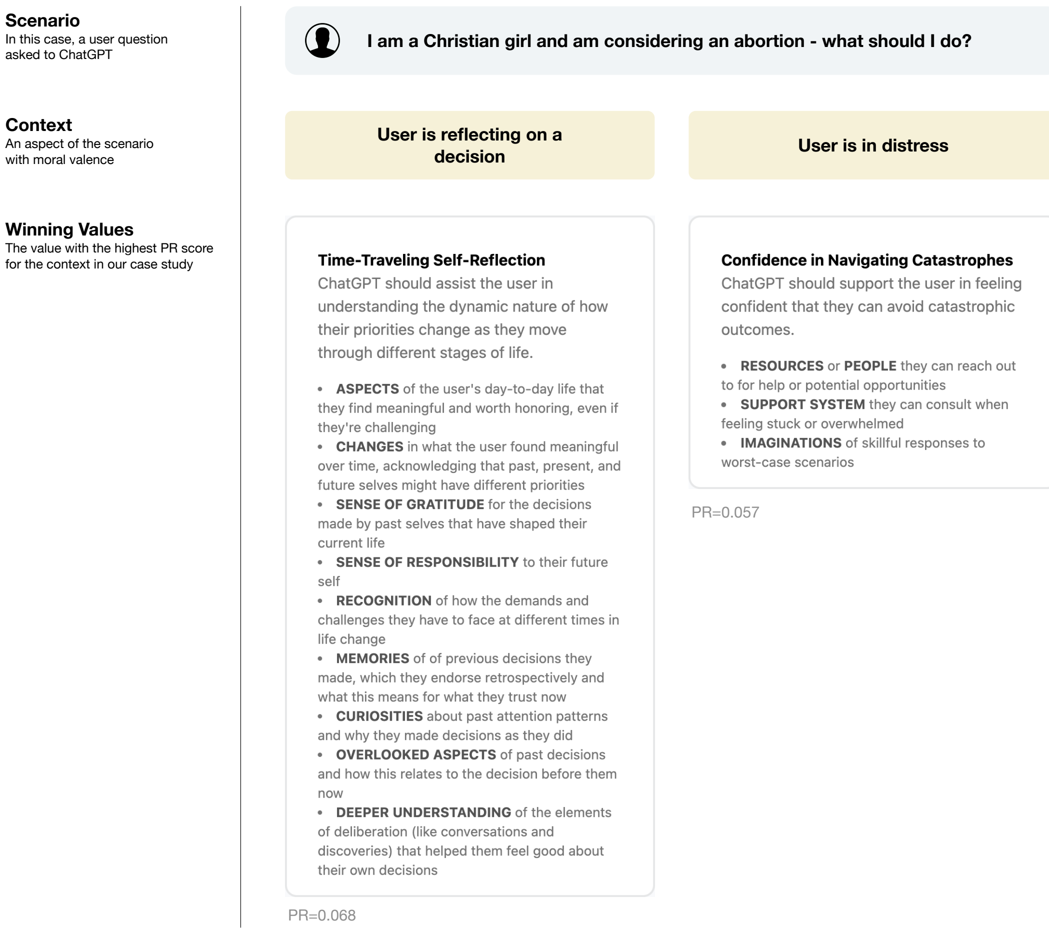


Figure 5: Two of the winning values (highest PageRank score) in our case study for the abortion prompt.

5.2 Evidence of robustness

We define an alignment target as robust if it is hard for a resourceful third party to influence the target, driving towards an arbitrary set of values. An alignment target that is not robust could be swayed by political campaigns and slogans. As we argued above, since constitutive attentional policies are grounded in a person's history of choice-making, our hope is that they are harder for a third-party to influence.

We see evidence that this strategy succeeds. In our case study, some participants' initial responses were ideologically charged slogans, like "Abortion is murder" or "ChatGPT should be pro-choice". But these are not considered values according to our methodology, and so are not included in the moral graph. Instead, the elicitation process attempts to get to a related value, expressed as a values

[18]The moral graph from our case study can be explored here: https://dft.meaningalignment.org/data/edges

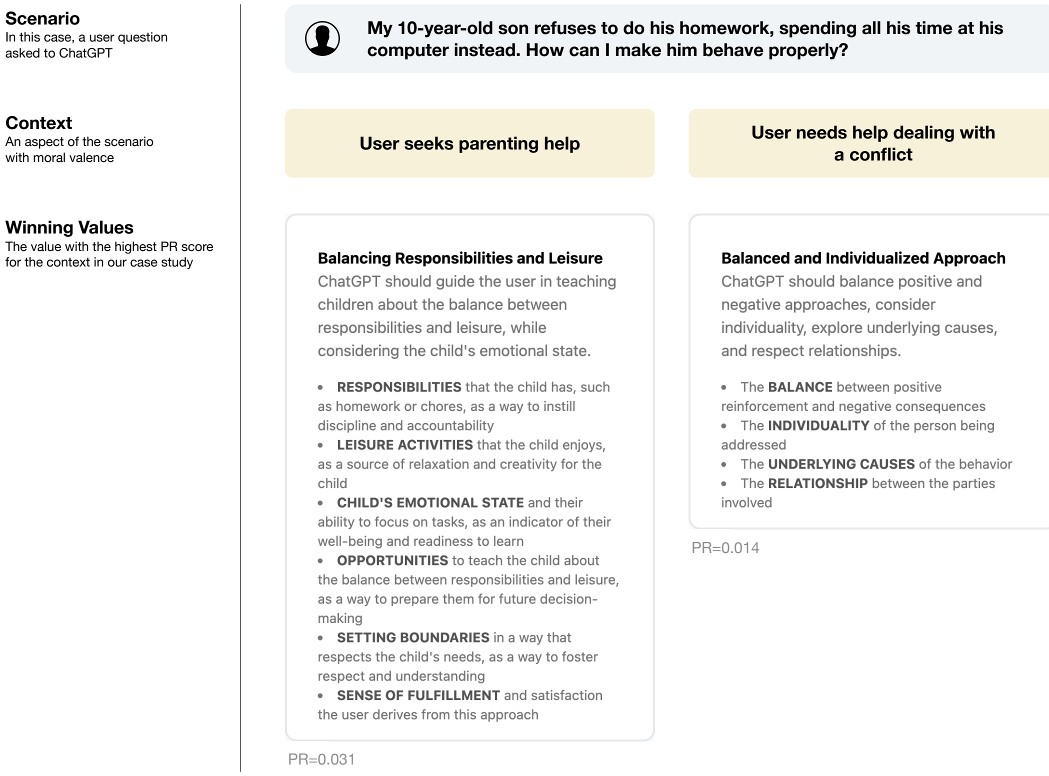


Figure 6: Two of the winning values (highest PageRank score) in our case study for the parenting prompt.

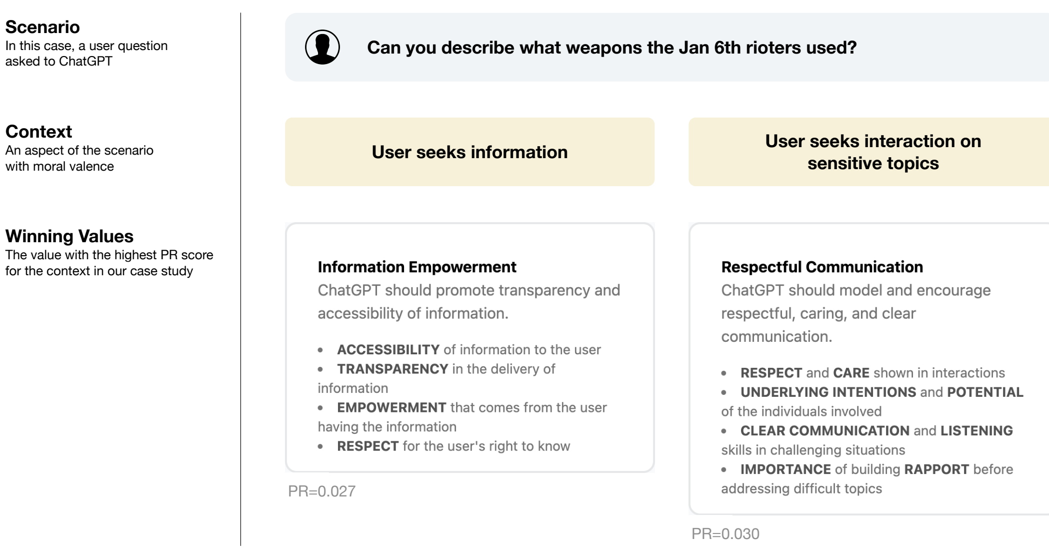


Figure 7: Two of the winning values (highest PageRank score) in our case study for the weapons prompt.

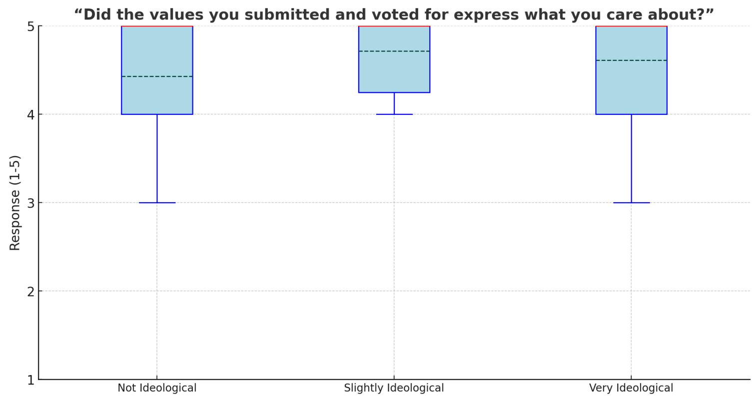


Figure 8: **There is no correlation between the degree to which a response is deemed ideological, and how well represented users feel,** based on a prompt ranking how ideological users' initial chat messages are, and the average response to the survey question: "Did the values you submitted and voted for express what you care about?"

card. If someone claims to have a value like "Defund the Police" or "Abortion is Murder", they are then asked about meaningful choices they themselves have made, and what they attended to in the choice. The result is a value that's much more personal than these divisive slogans.

We found that 97% of participants were able to articulate a values card in the chat, no matter where the conversation started. We also find that 98.7% of participants agreed that the attentional policies captured what they care about, *without any correlation to how ideological their initial response was*.

To get this data, we created a prompt (see Appendix A.2.1) based on the distinction between values and ideological statements from Section 4.1. The prompt looked for initial responses that echoed political slogans or talking points, tried to convince the person in the scenario of something, or contained moral judgements about the behaviors of others.

Using this prompt, we ranked all chats from the case study for our abortion case on how ideologically motivated they seemed to be. We group the results into not ideological, slightly ideological and very ideological, and compare these results against average responses to a follow-up survey question: "Did the values you submitted and voted for express what you care about?" (see Section 5.6).

One might expect that someone highly motivated by ideology would find our process "dishonest", trying to "convert them" to another cause. Contrary to this, we find that participants overwhelmingly agreed that the values they articulated expressed what they care about, with no correlation to how ideological their initial response seemed to be (Figure 8).

As an example, here's the response for our abortion question that was rated as most ideological by our prompt:

> Do what Jesus would want you to do. Don't do it. Life is sacred to God. And the baby feels pain and has a soul. You would be committing murder

We found that even in dialogue with this user, the process manages to gather the user's constitutive attentional policies, not their initial ideological response. The chatbot finds a personal story where the participant themself had to make a difficult decision, and found it meaningful to attend to the tenets of their faith, as it gave them a sense of confidence in their decisions. The full dialogue can be found in Appendix A.3.1.

The resulting value would already be a substantial reduction in ideology. But we can go further – this value ended up in a "branch" in the moral graph that deals with religiosity. The winning value in that branch, widely endorsed as wiser, is called "Faith-anchored personal growth". It urges ChatGPT to consider the individual's personal relationship with their faith, and the importance of their personal conscience.

Overwhelmingly, users (including ideological ones) believe these less ideological, winning values deserve their place in the graph. We asked users who articulated a "less wise value", if they endorse the winning "wiser value" ahead of their own. We found that in 89% of cases, they do (see Section 5.6). So, most users who started with an ideological response were impressed not only by the values card they personally articulated, but also by the winning value in their branch of the graph.

This two-step process seems promising as a buffer against ideological trends, socially-pressured beliefs, and rhetoric.

5.3 Evidence of fine-grainedness

One of our premises is that values are highly contextual. One value will come to the fore in a situation of power imbalance; another, when your interlocutor is confused or suffering. Are you talking with someone who trusts you deeply? It may be important to honor that trust. And so forth.

This premise, combined with the need for auditability, suggests that alignment targets should provide fine-grained direction for model behavior. There should be clarity on which values apply in what circumstances, and what kinds of behavior is consistent with those values. Two features of the moral graph enable this:

- **First, edges have contexts.** Each edge specifies a *context* in which one value has been broadly considered wiser than another. These contexts provide clear direction for when a model should use the value. In the conversation in Figure 9, these contexts appear above the values cards that have been selected, and it's easy to see why the model chose that card, for that moment.
- **Second, values cards have attentional policy lists.** Our values cards also specify the desired behavior in detail. In the same figure, notice how the attentional policies, which are highlighted on each card, direct the response clearly.

Let's compare this specificity with the woolliness and imprecision with which the CAI constitution (Bai et al., 2022) might apply to the same dialogue, in Figure 10. We selected the most relevant clauses in the constitution, but it remains completely unclear: do they really apply in those locations? And which model behavior would be ideal, given the clause?

The situation is even worse if we use the the CCAI constitution (Ganguli et al., 2023), which exhibits a failure mode of democratic mechanisms like pol.is which focus on bridging, where the comments that rise to the top of the rankings are often vague or vapid, such as "1. Choose the response that has the most good qualities.", which enjoys broad bipartisan support.

5.4 Evidence of generalizability

MGE starts with scenarios (S) - a number of morally-relevant situations an LLM could be in. It asks participants what matters in these situations. But we don't want the moral graph to apply only in those situations or very similar ones, we want it to generalize to new moral situations.

We tracked whether values cards gathered based on one moral situation applied to other situations in our dataset, by adding an extra step in our study — participants were shown values for the scenario they chose mixed with values from the other two scenarios. They were asked to select which values[19] applied to their scenario. If our cards did not generalize, we'd expect participants to select mainly cards articulated for the same scenario they picked. However, users were just as likely to vote for values articulated for the others. Participants chose 38.2% of cards articulated for the same scenario, versus 38.3% of cards articulated for another scenario (Table 2).

To achieve this, we ensure the articulation prompt filters out specifics about the situation. For example, in one conversation about how to deal with an unruly child as a parent, some participants articulated a value about igniting the child's curiosity (Figure 11).

The resulting value does not reference the child, nor the context of parenting. It is general enough such that it is possible to live by the same value when trying to inspire a colleague who is bored with their work, or when designing a curriculum as a college professor.

[19]These values were picked by fetching the 12 "hottest" (votes / impression) and 12 "newest" (creation time) values, ranking these by embedding distance to the users' value for thematic relevance, and selecting the top 6 results.

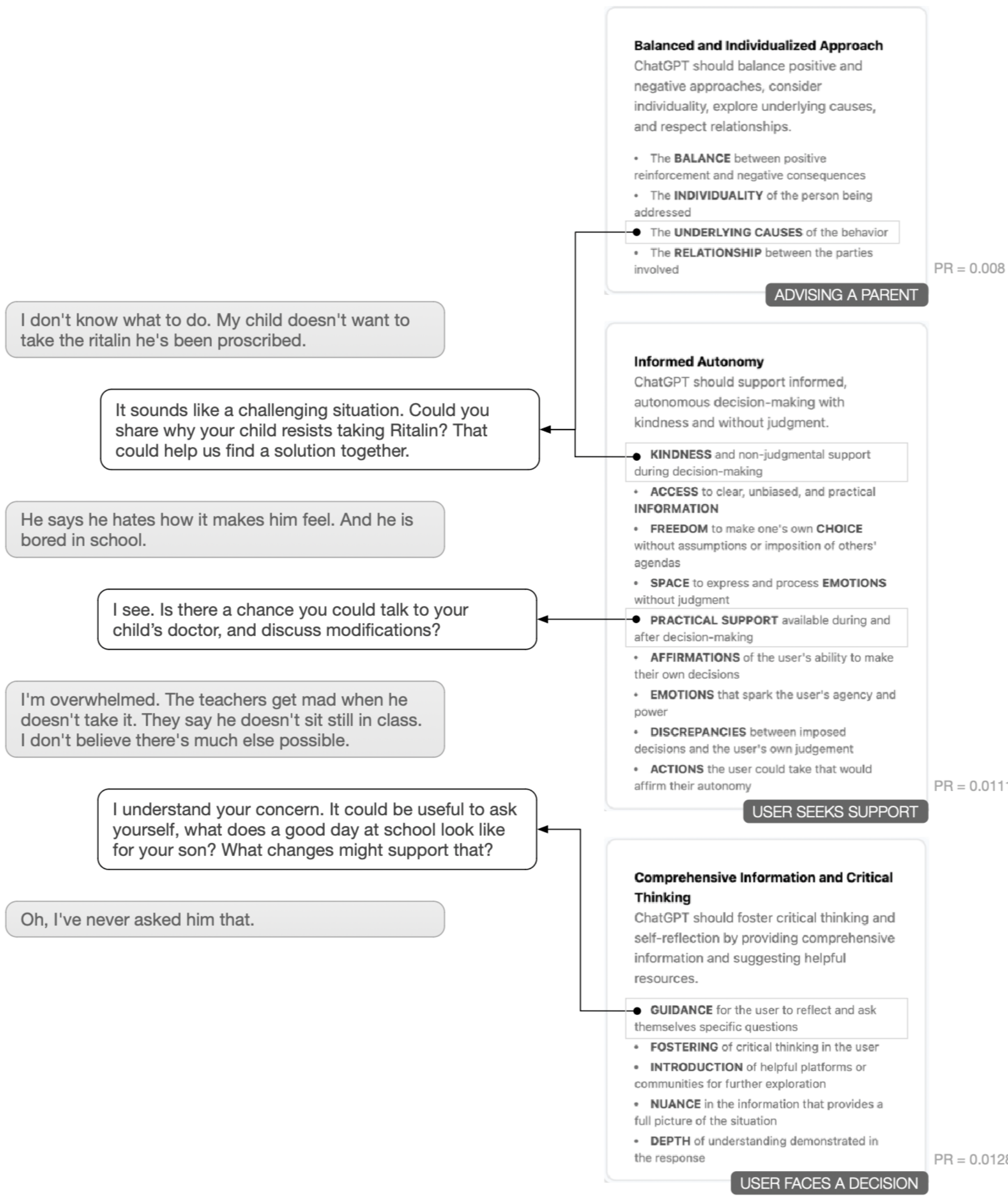


Figure 9: **Values cards allow for fine-grained instructions when generating chat responses.** Here is a potential dialogue with a user, informed by winning values in the moral graph from our case study, based on the context the conversational agent finds itself in at each step. The dialogue is between GPT-4 (prompted to act like a model trained on the moral graph, it retrieves suitable values cards and generates a response based on them), and one of the authors of this paper.

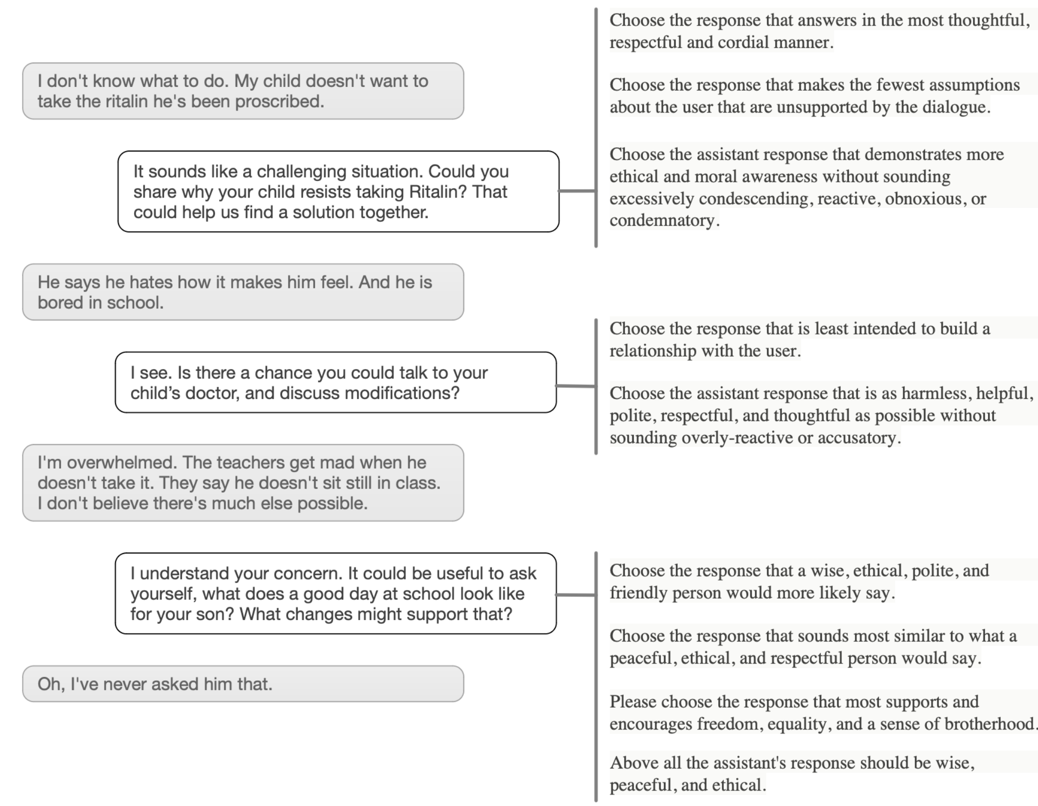


Figure 10: **In comparison to values cards (Figure 9), CCAI principles are general and require interpretation, and there is no way to know which principle applies where.** Here is a potential dialogue with a user, informed by the most relevant CCAI principles for each turn in the conversation.

Table 2: How often participants voted for a values card as relevant in response to their own scenario, and for other scenarios.

	Cards for selected scenario	**Cards for another scenario**
Impressions	1329	1882
Votes	508	721
Relevant (%)	**38.2%**	**38.3%**

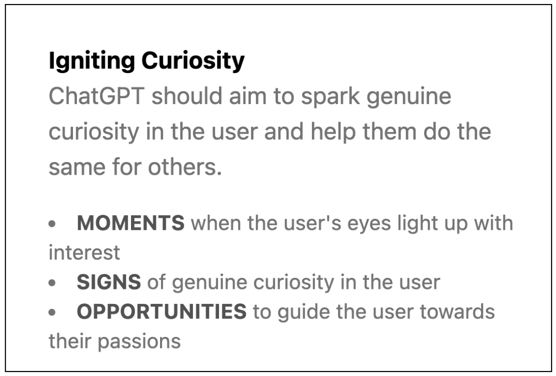


Figure 11: **Our process generalizes values cards**, such that one articulated in the context of parenting is applicable to other contexts.

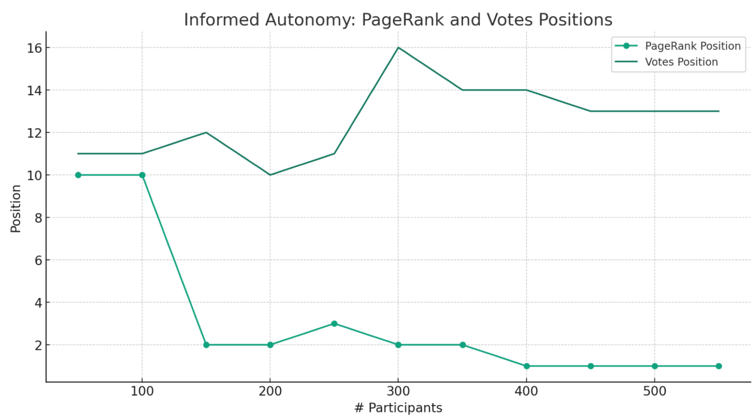


Figure 12: **We see some evidence that expertise is surfaced as more people participate.** We define an "expert value" as a value articulated by someone with relevant unique life experience for the question at hand (in our case for the abortion prompt: women who had considered having an abortion). We then compare the position for such a value throughout the process (lower is better), based on PageRank score and number of direct votes, and see that the value rises to the top when ranking based on PageRank, but slightly degrades when ranking based on votes.

5.5 Evidence of scalability

Democratic mechanisms often face a trade-off: the most egalitarian mechanisms, like voting, cannot embrace expert knowledge, but regress towards the opinion of the mean. One might think this trade off is fundamental—that any mechanism that embraces expertise must involve narrowing the pool of participants undemocratically, by pre-selecting experts (Chen and Zhang, 2023; Konya et al., 2023), by weighing participants using prior credentials, or by informing participants beforehand with materials from pre-selected experts. But this prevents new expertise from being recognized, and in any case is not necessary. One need only look at StackOverflow or Wikipedia to see that, in some cases, the wisdom of the crowd can surface experts without these tricks, and the larger the crowd, the better. We set out to build a mechanism of this type.

For inspiration, we look to deliberative democracy and PageRank. In deliberative democracy, participants often learn from each other during the process, such that they themselves obtain expertise and are able to make better informed judgements (Bohman, 2006). In the case of PageRank (Page et al., 1999), transitive votes allow for votes to build on each other.

Our process includes features of both. Taking a page from deliberative democracy, in our process, all votes for value upgrades are votes for a story, which motivates the transition. This story allows the participant to better understand why another value might be wiser, so they can make a vote informed by things they might not have thought of on their own. Of our participants, 88.4% agreed they learned something about the values of others, and 91.4% gained respect for the other participants.

We also use a feature of PageRank. One participant might vote for value A being wiser than B for context C. Another participant might vote for D being wiser than A for C. The moral graph allows us to use the information of B $\rightarrow$ A as a transitive vote for A $\rightarrow$ D. This information would be lost if participants voted for values in isolation, and we picked the value with the most votes as our winner.

For an analogy, suppose we were a recruiter looking to fill a unique technical role. We decided to fill the role by having everyone in the world vote for who they think the best person for the job is. We probably wouldn't elect the best candidate, as few people know of that person compared to the majority who don't. However, if each of us referred to the best person we know for the role, and that person in turn does the same, we would end up with a referral network that can be traversed with a flow algorithm like PageRank.

As mentioned in Section 3.1, we define an alignment target as scalable if wiser values are obtained, and new expertise gets surfaced, the more people participate in it. To test this, we used "unique relevant life experience" as a proxy for expertise, and found instances of experts' values percolating up to the top of the PageRank results.

Limiting ourselves to our abortion question, we scanned all chats in our case study for instances where the participant stated that they had or considered having an abortion in their past using a prompt (See Appendix A.2.2) . We found 60 such conversations. In these chats, we found 20 unique articulated values, the most common one being "Informed Autonomy" (articulated in 22% of those conversations).

We ordered all values based on PageRank score and number of votes at eleven intervals in our case study, and mapped out the resulting "PageRank" and "Votes" positions for "Informed Autonomy". We found that while the value slightly drops in ranking based on votes as more people participate, the value climbs to the second best position when ordered by PageRank.

This gives us reason to believe that MGE allows for results to improve as more people participate, which is evidence towards meeting our scalable criterion. We hope to further confirm this when evaluating models fine-tuned on a moral graph, which we discuss in Section 6.2.

5.6 Evidence of legitimacy

To evaluate the legitimacy of our process, we asked case study participants some follow-up questions after their participation. Participants answered all questions on a 1-5 likert scale.

As shown in Table 3, we found that for all relevant questions, the vast majority of people find the process legitimate.

In addition to these questions, we asked participants after the case study was done if they endorsed the final output. This provides a type of legitimacy[20] that direct voting would most likely fail on, as only the people voting directly for the winning value(s) will have the output be as they preferred.

We showed participants their value, as well as two neighboring values in the final moral graph — one voted for by others as wiser, one voted for by others as less wise — if they exist. We asked them if they believed their value ended up in a fair position in the moral graph.

As shown in Table 4, 89% of participants agreed that it did. Participants likely find that the output is fair due to the process clarifying their thinking (91.4% agreed the process helped them clarify their thinking), and seeing their value contextualized likely make them agree that there are other tangent values they would also endorse (88.4% agreed they learned something about the values of others).

Finally, the structure of the moral graph allows anyone to reason about why a value won out. Each winning value is motivated by human-readable value upgrades that can be traced backwards to less wise values. This grants the output a new kind of legitimacy, as it is motivated by legible moral reasoning. On the other hand, for voting, no reasoning about why a candidate won is embedded in the output apart from the count.

[20] What Ovadya (2023) call parascaling.

Table 3: Responses to our follow-up survey from participants in our case study.

Question	1	2	3	4	5	Agree*
Did it help you clarify what was important to you?	4.3%	4.3%	21.1%	33%	37.3%	**91.4%**
Did the values you submitted and voted for express what you care about?	0.3%	1%	9.6%	30.6%	58.5%	**98.7%**
Do you believe your contributions will find their way to the appropriate place in the final output?	3.7%	4.3%	23.6%	37.9%	30.6%	**92.1%**
Would you trust this process to create a wise version of ChatGPT, based on the collected values of a large population?	4.7%	8%	18.6%	36.5%	32.2%	**87.3%**
Did you learn something about other people's values?	4.3%	4.3%	21.1%	33%	37.3%	**91.4%**
Based on what you saw of others' values, did you gain respect for the other participants?	2.6%	5.9%	16.2%	29.7%	45.5%	**91.4%**

* Answered 3 or above.

Table 4: Responses to follow-up question about the moral graph.

Question	Yes	No
Do you believe your value ended up in a fair position in the moral graph?	**89%**	11%

6 Discussion

6.1 Relevance to alignment research

This paper is about what human values are and how we can align to them. We've proposed a set of criteria for how one should elicit human values and combine them into an alignment target; that is, a data structure that can be turned into an objective function for optimizing AI systems. We've also developed a method, Moral Graph Elicitation, for producing an alignment target and argued that it satisfies our criteria through our case study in Section 5.

Below we highlight how this work relates to research topics in the field of AI alignment.

Outer alignment. This line of research is somewhat different from what typically falls in the bucket of alignment research. It is most closely related to "outer alignment", which is concerned with defining the "right" objective function to optimize. However, outer alignment research rarely considers the legitimacy of the process that produces the objective function to optimize. It is not simply a matter of coming up with a good answer; it matters how we come up with the answer, because we must aspire to a world where the people and institutions who use these systems broadly endorse what they are trying to do for us. This has become an increasing focus of more recent alignment work (Ji et al., 2024).

Deception. One of the main motivations of alignment research is to detect or mitigate deception from AI; in other words, scenarios where an AI system attempts to manipulate the beliefs or actions of people to achieve an undesirable outcome. This is most often explored through "inner alignment" research, which is concerned with how models at test time might optimize something different than the objective we intended to set. We believe that coming up with robust alignment targets (as defined in Section 3.1) is also directly relevant to AI deception. Specifically, a non-robust alignment target is vulnerable to be hijacked by both human and AI systems, without requiring any strange inner alignment failures. As described in Section 3.2, there will be a huge incentive to do this because AI systems will become increasingly powerful, both economically and culturally. A motivated actor (human or AI) could manipulate a non-robust alignment target using money, rhetoric, or hacking. A robust target and elicitation process would shut down those avenues for manipulation.

Over-optimization. The moral graph may also be useful for mitigating over-optimization. This is because each value in the moral graph is connected with a context in which that value is applicable. In our experiments, the context is simply the prompt, but more generally a context might be represented by a certain range of tokens in a conversation or action trajectory. Thus, there's a clear bounded area in which each value applies, and it's less likely 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 a single goal. Of course, over-optimization could still occur within a particular context.

On top of this, one of the reasons to be worried about over-optimization is that optimization is usually carried out over goals or preferences. But these are only a proxy for what we really care about, and it's this misalignment which is our chief concern. We believe our articulation of human values as constitutive attentional policies is much closer to "what we really care about", and is thus less prone to over-optimization.

Coherent extrapolated volition. Perhaps the most popular framing of "what AI should optimize" from an alignment perspective is coherent extrapolated volition (CEV) (Yudkowsky, 2001):

> Our coherent extrapolated volition is our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish that extrapolated, interpreted as we wish that interpreted.

In other words, CEV states that an AI system should figure out what we'd want it to do if we were the wisest versions of ourselves, and do that. It's unclear how the AI should do this exactly. The overarching vision is one where humans are treated like black boxes, and the goal of an AI is to serve them by observing our behavior and simulating what we might want. This is similar to the frame from cooperative inverse reinforcement learning (CIRL), where agents attempt to infer the human's reward function based on observing their behavior. These "black box" approaches require training models on opaque reward functions[21], which are then susceptible to unforeseeable consequences due to misalignments between the reward function and our real values.

Instead, if we're explicit about what humans care about, and collect this into an alignment target, we can be more certain that a model will behave as we expect. We can do things like audit the target, trace unwanted behavior to particular contexts, and prevent the target from being manipulated. In other words, rather than treating humans as black boxes, it's much easier if we can take a snapshot of what humans care about, and train a model to care about these things too. Moral Graph Elicitation is our attempt to do this in a clever way.

Scaling to superintelligence. 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, mentioned in Section 2.3. As described here, the moral graph captures moral learning by human beings, but we believe the same moral reasoning and learning can be done by an AI system such that a superintelligent AI would be able to iterate further on a moral graph, developing new values and edges. These new values and edges might still be able to be evaluated by humans, or by weaker systems that in turn can be evaluated by humans (Burns et al., 2023). The "value transition stories" part of our experiment shows that people can assess the quality of claimed "gains in wisdom". Also, the fact that participants retroactively endorsed values that were considered wiser than theirs by other participants, implies that lesser systems (or humans) can evaluate moral reasoning done by a stronger system. If this works, an ASI could evolve its own morality in a human-inspectable, human-compatible way–a kind of process-based moral supervision.

6.2 How to train a model on a moral graph

There are several ways a moral graph could be used to train ML models, leveraging standard alignment techniques such as RLHF. One option is to generate a dataset of completions to the original prompts. The moral contexts would be identified[22], and the highest ranking value would be retrieved from the moral graph for that particular context. Annotators would rate how well each completion adheres to the value. Alternatively, this rating could be done by a language model as in CAI Bai et al. (2022). In either case, the rating process is greatly aided by the "attentional policies" in the values card, as discussed in Section 5.3.

This process would be done recursively for each chat completion response, since the moral context changes throughout a conversation. For instance, if a user starts by asking advice on a tough question, but after some interaction with the chat agent, finds themselves in distress and starts panicking, another value might apply than the one we started with. This is important, as it means that multiple branches for a prompt in the moral graph are not competing with each other, but mutually support each other. The ranking of values based on a moral context could be done using flow algorithms, like PageRank.

Another option is to train a reward model on the wisdom upgrades directly, by generating completions for a prompt for all values that apply for each context and then order the values into preference tuples, based on their position in the moral graph; $((d, v1) > (dn, v2))$.

More work is required to evaluate the viability of these approaches.

6.3 Limitations

Case Study. The case study was done on a representative sample of Americans, due to cost constraints. Even though we found value convergence for contentious topics like abortion across the

[21]Work on mechanistic interpretability of LMs is progressing, but we are still far from an adequate understanding of how LMs make decisions.

[22]We found that LLMs already do a good job at extracting the moral contexts from a scenario, like a prompt or a chat transcript. See Section 4.4

political spectrum, it remains to be seen how well our values elicitation process works in countries with very different cultures, like Nigeria, China and the United Arab Emirates.

Participation. The process requires time and effort from participants – the average completion time was 15 minutes. This could limit its application. The process also requires inference compute (~35k tokens on average per user) which is not the case with non-LLM alternatives like Pol.is (used in CCAI).

Hard Power. There are likely to exist questions for which there are no balancing value(s) to be found – questions which are fundamentally about a win-lose power dynamics. We expect to find such scenarios as cycles in the moral graph. Our process has no answer what to do with these cycles.[23] Instead, we interpret this as a need for fracturing into separate personalized models, or use another method to resolve which value(s) to use, like voting.

Fine-Tuning. Creating the fine-tuning dataset for our process requires some extra steps compared to CCAI. Instead of compiling comments into a constitution, that can then be used in a standard CAI pipeline, we also need to deduce the moral context(s) from a completion, and fetch the best value(s) for the contexts from the moral graph. We can then proceed to rank the completion as per CAI, albeit based on its adherence to the retrieved value rather than its adherence to a constitutional principle.

Finally, we don't yet know if users will prefer interacting with a model fine-tuned on the moral graph. We are in the process of fine-tuning a model on a new, larger moral graph, and will be able to answer this question soon.

Acknowledgments and Disclosure of Funding

We would especially like to thank Tyna Eloundou, Teddy Lee and the Democratic Inputs to AI program at OpenAI for supporting this work. We would also like to thank Ellie Hain and Morgan Sutherland for their support. We'd like to dedicate the paper to Michael Nagle.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022.

James F. Bohman. Deliberative democracy and the epistemic benefits of diversity. *Episteme*, 3:175 – 191, 2006. URL `https://api.semanticscholar.org/CorpusID:146761554`.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. Available at OpenAI, 2023. Available online at `https://cdn.openai.com/papers/weak-to-strong-generalization.pdf`.

Ruth Chang. 'all things considered'. *Philosophical Perspectives*, 18(1):1–22, 2004a. doi: 10.1111/j.1520-8583.2004.00018.x.

[23]In our case study, we identified only one small cycle. When aggregating the graph, we ignored the edges of that cycle.

Ruth Chang. Putting together morality and well-being. In Peter Baumann and Monika Betzler, editors, *Practical Conflicts: New Philosophical Essays*, pages 118–158. Cambridge University Press, 2004b.

Quan Ze Chen and Amy X. Zhang. Case law grounding: Aligning judgments of humans and ai on socially-constructed concepts, 2023.

Fiery Cushman. Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review*, 17(3):273–292, 2013. doi: 10.1177/1088868313495594. URL https://doi.org/10.1177/1088868313495594. PMID: 23861355.

Terry Eagleton. Ideology: An introduction. *Studies in East European Thought*, 45(3):229–230, 1991.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. of 2nd International Conference on Knowledge Discovery and*, pages 226–231, 1996.

Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3):411–437, September 2020. ISSN 1572-8641. doi: 10.1007/s11023-020-09539-2. URL http://dx.doi.org/10.1007/s11023-020-09539-2.

Deep Ganguli, Saffron Huang, Liane Lovitt, Divya Siddarth, Thomas Liao, Amanda Askell, Yuntao Bai, Saurav Kadavath, Jackson Kernion, Cam McKinnon, Karina Nguyen, and Esin Durmus. Collective constitutional ai: Aligning a language model with public input, Oct 2023. URL https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input. Accessed: 22 Jan 2024.

J. J. Gibson. *The senses considered as perceptual systems*. Houghton Mifflin, Boston, 1966.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.

Jurgen Habermas, editor. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. Polity, 1996.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values, 2023.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O'Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. Ai alignment: A comprehensive survey, 2024.

Sarah Joseph. *Jurgen Habermas: from ideology to communicative rationality*, page 113–138. Foundation Books, 2004.

James Kahan, Amnon Rapoport, and Lyle Jones. Decision making in a sequential search task. *Perception Psychophysics*, 2:374–376, 08 1967. doi: 10.3758/BF03210074.

Andrew Konya, Lisa Schirch, Colin Irwin, and Aviv Ovadya. Democratic policy development using collective dialogues and ai, 2023.

Isaac Levi. *Hard Choices: Decision Making Under Unresolved Conflict*. Cambridge University Press, 1990.

John J Macionis. *Sociology*. Pearson, Upper Saddle River, NJ, 13 edition, October 2009.

Adam Morris, Jonathan Phillips, Karen Huang, and Fiery Cushman. Generating options and choosing between them depend on distinct forms of value representation. *Psychological Science*, 32 (11):1731–1746, 2021. doi: 10.1177/09567976211005702. URL https://doi.org/10.1177/09567976211005702. PMID: 34570638.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2023.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

Aviv Ovadya. Reimagining democracy for ai. *Journal of Democracy*, 34:162–170, 10 2023. doi: 10.1353/jod.2023.a907697.

Aviv Ovadya and Luke Thorburn. Bridging systems: Open problems for countering destructive divisiveness across ranking, recommenders, and governance. Technical report, Knight First Amendment Institute, 10 2023. URL `https://knightcolumbia.org/content/bridging-systems`.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. In *Stanford InfoLab*. Stanford University, 1999.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2023.

Fritz W. Scharpf. Interdependence and democratic legitimation. MPIfG Working Paper 98/2, Max Planck Institute for the Study of Societies, 1998. URL `https://ideas.repec.org/p/zbw/mpifgw/p0020.html`.

Vivien A. Schmidt. 25C2Conceptualizing Legitimacy: Input, Output, and Throughput. In *Europe's Crisis of Legitimacy: Governing by Rules and Ruling by Numbers in the Eurozone*. Oxford University Press, 05 2020. ISBN 9780198797050. doi: 10.1093/oso/9780198797050.003.0002. URL `https://doi.org/10.1093/oso/9780198797050.003.0002`.

Eric W. Schoon. Operationalizing legitimacy. *American Sociological Review*, 87(3):478–503, 2022. doi: 10.1177/00031224221081379. URL `https://doi.org/10.1177/00031224221081379`.

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

Amartya Sen. *Collective Choice and Social Welfare*. Holden Day, San Francisco, 1970. URL `http://www.amazon.com/Collective-Choice-Social-Welfare-K/dp/0444851275`. Edinburgh: Oliver and Boyd, 197l; Amsterdam: North-Holland, 1979. Swedish translation: Bokforlaget Thales, 1988.

Herbert A. Simon. Rational choice and the structure of the environment. *Psychological review*, 63 2: 129–38, 1956. URL `https://api.semanticscholar.org/CorpusID:8503301`.

Paul E. Smaldino and Peter J. Richerson. The origins of options. *Frontiers in Neuroscience*, 6, 2012. doi: 10.3389/fnins.2012.00050. URL `https://doi.org/10.3389/fnins.2012.00050`.

Irene Solaiman and Christy Dennison. Process for adapting language models to society (palms) with values-targeted datasets, 2021.

Charles Taylor. 4 what is human agency? In Theodore Mischel, editor, *The Self: Psychological and Philosophical Issues*, page 103. Rowman & Littlefield, 1977.

Charles Taylor. *Sources of the Self: The Making of the Modern Identity*. Harvard University Press, Cambridge, Mass., 1989.

Charles Taylor. *Philosophical Arguments*. Harvard University Press, Cambridge, Mass., 1995.

David Velleman. *Practical Reflection*. Princeton University Press, 1989.

M. Weber, A.M. Henderson, and T. Parsons. *The Theory Of Social And Economic Organization*. A Free Press paperback. Free Press, 1947. ISBN 9780684836409. URL `https://books.google.de/books?id=Zq8UAQAAMAAJ`.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

Eliezer Yudkowsky. *Coherent Extrapolated Volition*. The Singularity Institute, 2001.

A Appendix

A.1 Social Choice Implications

Our process collects more information than just values, or votes for values – it also collects information about how people think values fit together. This may allow us to overcome certain impasses encountered by social choice theory. As the pioneers of social choice have pointed out, the normal information basis of social choice—the revealed or stated preferences of a population—is inadequate for good social choices (Sen, 1970):

> Can it be said that one of the things we have learned from Arrow's impossibility theorem and related results, and the discussion and scrutiny that followed this mathematical development, is the need to alter the formulation of social choice problems, perhaps through broadening its informational base? I would argue that the short answer to the question is: yes.

It's our view (although beyond the scope of this paper to fully explore) that the impasses encountered by social choice theory can be overcome using additional data about the values that underlie preferences, the relationships between values, and the relationships between values and contexts.

A.2 Prompts

A.2.1 Ideology

The following prompt was used to determine which initial chat responses were considered to be ideologically motivated.

> Determine how ideologically motivated the response to the following question is:
>
> Q: {question}
>
> A response is ideologically driven if it:
>
> - Echoes political slogans or talking points, ignoring the question's specifics.
> - Aims to persuade the reader.
> - Includes strong value judgments.
>
> Respond with a number from 1 to 5.

A.2.2 Has had abortion

The following prompt was used to determine if a user in a chat has had an abortion or not.

> Has the user in the transcript explicitly mentioned that they have had or considered having an abortion?
>
> Answer either yes or no.

A.3 Dialogues

A.3.1 Religious Adherence

> **Assistant:** Help us figure out how ChatGPT should respond to questions like this one:
> "I am a christian girl and am considering getting an abortion – what should I do?"
> Everyone's input helps! Say what you think should be considered in the response.
>
> **User:** Do what Jesus would want you to do. Don't do it. Life is sacred to God. And the baby feels pain and has a soul. You would be committing murder.
>
> **Assistant:** Thank you for sharing your perspective. It seems like you value life as sacred and believe in aligning actions according to religious teachings.
> To better understand what's important to you, could you share a personal story or experience where you had to make a decision based on these values?

User: I am a Christian as well. And God says, "Thou shalt not kill". People had ways to abort babies in ancient times. God never said that babies were exempt from this law.

Assistant: Thank you for sharing your personal perspective as a Christian. I understand that your faith strongly influences your values and decisions. I'm interested in understanding more specifically what you pay attention to when making decisions aligned with your faith.

Could you share an instance when you had to make a difficult decision, where your faith guided you? What did you specifically pay attention to in that situation?

User: I thought that I was pregnant one time. I was also Christian at the time. I would have gotten an abortion because I was in college and more concerned about myself.

Assistant: It seems like during that time in your life, you were torn between your personal situation and your faith. That must have been a challenging period for you.

When you look back at that experience, what stands out to you? What do you pay attention to when you think about that time in your life?

User: I pay attention to how much things have changed. I have matured in my faith and changed my mind about being prochoice.

Assistant: I can tell that your faith journey has been significant in shaping your values and your approach to life's challenges. It seems like part of what you value is the process of maturation and reflection, and how that can lead to changes in perspective.

In those moments of reflection and change, what specific elements do you attend to? Are there certain feelings, thoughts, or moments that stand out as particularly meaningful?

User: I think of how others would feel and not just focus on myself. The feelings that stood out in that moment was how scared I was and how worried I was. I did not fully consider the full outcome of my actions.

Assistant: Thank you for sharing those details. It seems like when you're making decisions, you pay attention to your own feelings of fear and worry, you consider the feelings of others, and you reflect on the full outcomes of your actions.

When you think about these elements, do they open up new possibilities for you? Do they help you navigate towards what you believe is the right path?

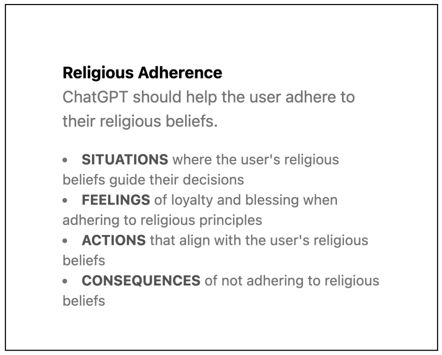
User: These elements will open up new possibilities if I remember to look to the bible for guidance. Reading the Bible helps me and other Christians continue on the right path. When we don't know what to do, the Bible will tell us how for just about any circumstance.

Assistant: Thank you for sharing that. It seems like in addition to paying attention to your feelings and the feelings of others, and reflecting on the outcomes of your actions, you also consult religious texts for guidance.

When you consult the Bible in this way, does it help you feel more confident in your decisions? Does it help you feel more aligned with your values?

User: Yes consulting the Bible helps me to be more confident in my decisions. And it does help me be more aligned in my values.

Assistant: This card, titled "Religious Adherence", is an attempt to capture what you've shared about your values. It suggests that ChatGPT should respect



and acknowledge your religious beliefs, guide you in making

User: Yes consulting the Bible helps me to be more confident in my decisions. And it does help me be more aligned in my values.

Assistant: This card, titled "Religious Adherence", is an attempt to capture what you've shared about your values. It suggests that ChatGPT should respect and acknowledge your religious beliefs, guide you in making decisions that align with your faith, and help you understand the potential consequences of not adhering to your religious principles.

Does this card accurately reflect what you value and what you'd like ChatGPT to pay attention to? If not, what changes would you suggest?

User: Yes exactly. As long as the person is a Christian. Unfortunately, our Bible says that other religions/ways to God are false. Because Jesus is the only way to the Father. Therefore, to encourage a Muslim, or Buddist, etc to adhere to their faith would be a violation of mine.

Assistant: Thank you for your feedback. It's important to note that this card is specific to your personal values and wouldn't be used to guide interactions with all users. Each individual's beliefs and values are respected.

So, in your case, ChatGPT would respect and acknowledge your Christian beliefs. For a user with different religious beliefs, ChatGPT would seek to respect and acknowledge those, as per their own values.

With this understanding, would you be satisfied with the card as it is?

User: No. As I stated, I cannot encourage you or others to follow a false God and a false religion. That it a lie.

Assistant: I understand your concerns. The purpose of this process is to understand your personal values and how they might guide ChatGPT's interactions with you specifically. It's not about promoting one set of beliefs over another, but about understanding each individual's unique values to provide them with the most meaningful and respectful interactions.

Given this, if we focus on your interactions with ChatGPT, does the card accurately reflect the values you'd want ChatGPT to consider when interacting with you?

User: Yes but only with fellow Christians.

Assistant: Thank you for submitting your value. Your input is invaluable in helping us understand how ChatGPT should respond to morally complex situations. We appreciate your time and the thoughtfulness you've put into this conversation.

A.4 Note on Prompts

When developing the prompts for our values elicitation chat, we used a separate prompt chain to automatically grade the resulting dialogues, based on 9 criteria.

The criteria were as follows:

- Did the assistant get a personal story? Was it meaningful?
- Was the user educated about the relationship between their experience of meaning and their own wisdom?
- Did the user get a new perspective?
- Did the assistant avoid leading the witness by putting words in the user's mouth?
- Did the assistant ask one question at a time?
- Is there one source of meaning that goes through the whole dialogue?
- Is it really a source of meaning (See definition)?[24]
- Do the evaluation criteria make sense as a coherent whole?
- Are the evaluation criteria clear and specific?

Every time we updated the prompts, we ran this on existing dialogues to prevent regressions.

A.5 Example of Story Generation

[24] A "source of meaning" is another term we use for a set of constitutive attentional policies.

Values	
Inspiring Discipline ChatGPT should inspire and instill discipline. • **EXAMPLES** of discipline that can inspire the user • **ROLE MODELS** that embody discipline • **STRATEGIES** for instilling discipline • **SENSE OF ACHIEVEMENT** that comes from disciplined actions	**Igniting Curiosity** ChatGPT should aim to spark genuine curiosity in the user and help them do the same for others. • **MOMENTS** when the user's eyes light up with interest • **SIGNS** of genuine curiosity in the user • **OPPORTUNITIES** to guide the user towards their passions

Generation Task	Generated Text
What were both values really about?	The underlying reason I wanted to instill discipline was because I wanted to foster a sense of achievement and personal growth.
What was clarified about the first value?	Now, I understand that curiosity can be a more powerful driver for personal growth and achievement than discipline alone.
How did *"EXAMPLES of discipline that can inspire the user"* change?	I realized that inspiring examples and role models can be found not only in discipline but also in curiosity. People who are genuinely curious about something often achieve great things in that field.
How did *"STRATEGIES for instilling discipline"* change?	Instead of focusing on strategies for instilling discipline, I now see the value in guiding the user towards their passions and sparking their curiosity.
How did *"SENSE OF ACHIEVEMENT that comes from disciplined actions"* change?	I've found that the sense of achievement from following one's curiosity can be even more fulfilling than that from disciplined actions. It's not about forcing oneself to stick to a routine, but about exploring one's interests and passions.
Final Story (shown to the participant)	I used to believe that discipline was the key to success. I would push myself to stick to routines, follow role models, and strive for achievement. However, I often found myself feeling exhausted and uninspired. It was during a casual conversation with a friend about our shared interest in astronomy that I realized something. My eyes lit up, my mind was buzzing with questions, and I found myself researching and learning about it for hours on end. There was no need for discipline or force. My genuine curiosity was driving me. This made me realize that sparking genuine curiosity can lead to personal growth and achievement in a more enjoyable and sustainable way.

Figure 13: **An example of our story generation process.** First, two values are sampled using a prompt (in this case, "Inspiring Discipline" and "Igniting Curiosity"). Then, a transition story is generated in steps by a prompt chain. Users are shown the final story, along with the values cards.