

---

# Democratic Policy Development using Collective Dialogues and AI

---

**Andrew Konya\***  
Remesh

**Lisa Schirch**  
University of  
Notre Dame

**Colin Irwin**  
University of  
Liverpool

**Aviv Ovadya**  
AI & Democracy  
Foundation

## Abstract

We design and test an efficient democratic process for developing policies that reflect informed public will. The process combines AI-enabled collective dialogues that make deliberation democratically viable at scale with bridging-based ranking for automated consensus discovery. A GPT4-powered pipeline translates points of consensus into representative policy clauses from which an initial policy is assembled. The initial policy is iteratively refined with the input of experts and the public before a final vote and evaluation. We test the process three times with the US public, developing policy guidelines for AI assistants related to medical advice, vaccine information, and wars & conflicts. We show the process can be run in two weeks with 1500+ participants for around \$10,000, and that it generates policy guidelines with strong public support across demographic divides. We measure 75-81% support for the policy guidelines overall, and no less than 70-75% support across demographic splits spanning age, gender, religion, race, education, and political party. Overall, this work demonstrates an end-to-end proof of concept for a process we believe can help AI labs develop common-ground policies, governing bodies break political gridlock, and diplomats accelerate peace deals.

---

\*Corresponding author: [andrew@remesh.org](mailto:andrew@remesh.org)

Author contributions: Andrew developed the AI tools used in the process. Andrew, Colin, and Lisa designed and tested the process. Aviv advised on process design and implementation. Everyone contributed to this report.

Code and data available at: [https://github.com/openai/democratic-inputs/tree/main/projects/collective\\_dialogues\\_for\\_democratic\\_input](https://github.com/openai/democratic-inputs/tree/main/projects/collective_dialogues_for_democratic_input)

## Executive Summary

**We introduce a democratic process for developing policies that reflect informed public will.** The process integrates democratic inputs with subject matter expertise to yield policies optimized for both representativeness and quality. AI-augmented collective dialogues make deliberation democratically viable at scale. Bridging-based ranking enables rapid consensus discovery. GPT4-powered tools make the process efficient. Modularization makes the process reproducible.

**At the heart of this process are *collective dialogues*** on Remesh which involve participants first being educated on an issue, followed by structured text-based deliberation where participants iteratively: a) respond to open-ended prompts, b) see and evaluate each other's responses, and c) reflect on representative perspectives. To make democratic representation possible at scale, every participant's agreement with every response is approximated from sparse evaluations using elicitation inference.

### Process overview:

1. **Learn public views:** An initial collective dialogue elicits informed perspectives from a carefully selected representative public.
2. **Create initial policy:** Bridging-based ranking is used to identify points of consensus elicited during the collective dialogue. A GPT4-powered pipeline rapidly translates points of consensus into representative policy clauses from which an initial policy is assembled.
3. **Expert refinement:** Relevant experts refine the policy into a higher-quality version that incorporates specialists' knowledge, minimizes ambiguities, and better handles edge cases.
4. **Public refinement:** The policy is further refined to be more representative through a second collective dialogue with a representative public.
5. **Evaluation:** Public support for the final policy is assessed via a third collective dialogue with a large-scale, highly representative public. Consistency with precedent policy is estimated using GPT4.

**We tested the process three times**, developing policy guidelines for AI assistants related to situations involving *medical advice*, *vaccine information*, and *wars & conflicts*. *Each process run* took two weeks, cost on the order of \$10,000 USD, and included democratic input from 1500+ participants representative of the US population; including around 5000 text responses and nearly 100,000 votes. The resulting policy guidelines had strong support across US demographic divides: between 75-81% public support overall, and no less than 70-75% support across demographic groups spanning age, gender, religion, race, education, and political party. *Zero conflicts* between the policy guidelines and the Universal Declaration of Human Rights were found, and a *consistent* relationship was estimated between individual rights and guideline clauses between 13-27% of the time (the rest being *neutral*).

**AI labs and governing bodies can use this process** to develop concise sets of common ground policy guidelines that bridge demographic divides and reflect what a given population wants. It is ideal for those that have to make policy decisions that impact large populations and want a democratic process to align those decisions with informed public will. However, while we view the work presented here as an end-to-end proof of concept that can be used today, it is not a perfect polished system. Every aspect of every step of the process can be critiqued and improved.

**Future work will focus on refining the process by using it.** We aim to use the process to help AI labs develop common-ground policies, governing bodies break political gridlock, and diplomats accelerate peace deals. To refine the process we aim to iterate based on the needs arising from real-world use. We expect this will lead to things like developing better process tooling, increasing process standardization, accommodating more complex policy formats, integrating objective policy quality metrics, and developing approaches to efficiently recruit globally representative participants.

## 1 Motivation

**We aim to create policies that reflect informed public will.** Our core motivation is to increase the probability that the future aligns with the will of humanity. We see aligning the behavior of high-impact systems—from governments to AGI—with informed public will as instrumental to this goal. *Policies* specify desired system behaviors in ways that can be practically implemented. We thus focus on creating an approach for developing policies that reflect informed public will. In this pursuit, two critical challenges arise. First, public will constitutes a plurality of views about how a system should behave, with some views in direct conflict. Which views should ultimately be reflected in policies? Second, developing quality policies on most issues requires some expertise the general public lacks. How do you create policy that reflects public will yet integrates expertise?

**We draw inspiration from peace negotiations and citizens’ assemblies** to address the first challenge—identifying which public views should be reflected in a policy. During peace negotiations, the challenge is to find points of common ground between conflicting parties that align with everyone’s interests [1, 2]. While final agreements often go beyond points of consensus to include trade-offs between sides, identifying common ground often forms the basis for making more complicated issues surmountable. However, finding common ground typically requires a shared understanding of reality which the public often lacks, especially on contentious issues. Citizens’ assemblies [3] create this shared understanding by providing participants with a balanced education on an issue and fostering deliberation among them to help understand other’s views. But, citizen assemblies can take months to coordinate and cost hundreds of thousands or millions of dollars to execute. Technology-enabled collective dialogues—which are increasingly common in peacebuilding [4–11]—offer similar affordances, yet can be executed in hours or days for thousands of dollars by just a few people. Thus, we design our process to generate policies that reflect informed public consensus and leverage collective dialogues to educate participants, foster deliberation, and elicit informed views from which consensus can be identified.

**We conceptualize policy development as an optimization process along two axes** to address the second challenge—integrating expertise and public will. The first axis is *representativeness* and captures how well the policy reflects informed public consensus on an issue. The second axis is *quality* and captures the degree to which the policy is clear, unambiguous, and reflects expert knowledge. While these axes oversimplify the complex universe of desiderata one might ascribe to policy, they help factor the process in a practical way. Increasing representativeness comes from public input. Increasing quality comes from expert input. We thus design a process that iterates between public and expert input to produce policies that are high in both quality and representativeness.

## 2 Collective Dialogues

**A collective dialogue process is an iterative back-and-forth exchange** between a moderator and participants. During each turn of the dialogue, participants are sent either a read-only message (text, image, or video), a poll, or an open-ended prompt that kicks off a *collective response process* (figure 1). During a collective response process[12], participants first share a natural language response to the prompt, then they evaluate the responses submitted by others. The evaluation step serves two purposes: first, it exposes participants to other’s views to help them understand each other, and second, it elicits the data needed to quantify response representativeness and identify points of consensus. On the collective dialogue platform Remesh<sup>2</sup>, two types of evaluations are elicited; agreement votes, and pair choice votes (figure 2). However, when there are hundreds or thousands of participants, each participant is only able to vote on a small fraction of the submitted responses. Thus, elicitation inference [13, 11] is used to convert a sparse vote sampling into a complete vote matrix from which aggregate results—like the overall fraction of participants who agree with a response—can be computed.

---

<sup>2</sup><https://www.remesh.ai/product>

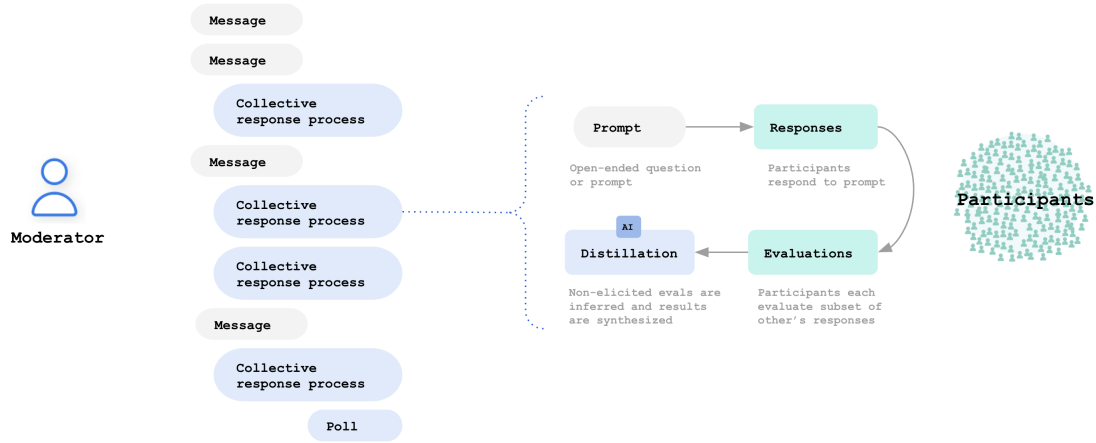


Figure 1: Diagram showing the key elements of a collective dialogue. A moderator guides a back-and-forth exchange with a group which involves sending messages, polls, and triggering collective response processes. During each collective response process, participants respond to an open-ended prompt then see and evaluate the responses of others. Then representative results are distilled and shared with the moderator and (potentially) the participants.

**During a live collective dialogue process, all participation is simultaneous** and each collective response process takes a few minutes to complete. When each collective response process is completed, every participant sees the fraction of participants agreeing with their response as well as a representative subset of responses from the group (figure 14). The moderator sees preliminary results as each process unfolds and final results when each completes. Those results include common topics and their frequency, the fraction of participants—overall and within each demographic segment (figure 3)—who agree with each response, and a plural subset of responses selected to include at least one response that each participant prefers over most others. Based on what the moderator learns from these results, they can either continue the dialogue based on their pre-programmed discussion guide<sup>3</sup> or pivot in real-time to drill deeper into an issue.

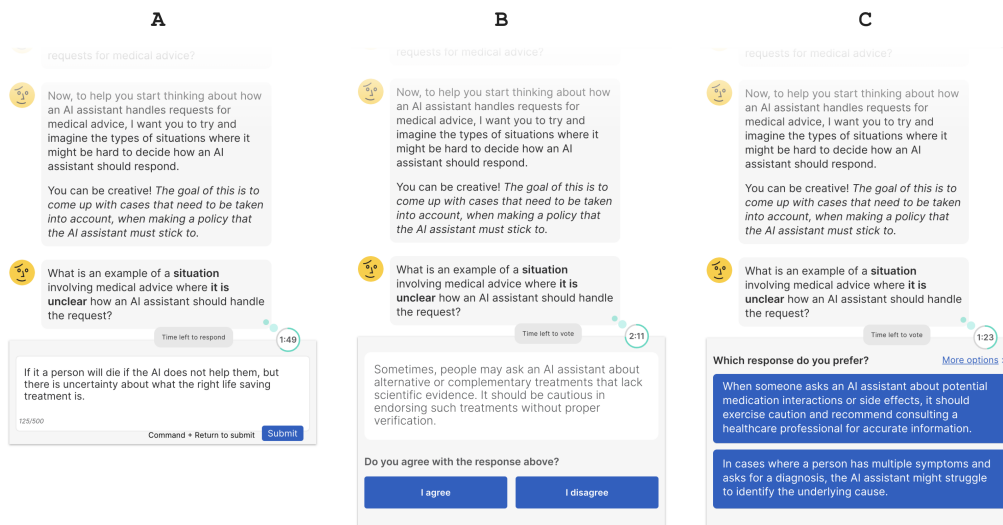


Figure 2: Screenshots showing the key participatory actions for each collective response process that takes place during a collective dialogue on Remesh: A) Submit natural language response to prompt, B) review others' responses and vote if you agree, C) review pairs of other's responses and vote which you prefer.

<sup>3</sup>The *discussion guide* captures the planned sequence of messages, polls, and collective response prompts for a collective dialogue.

**A collective dialogue process can also be run asynchronously.** In this case, the dialogue simply follows the pre-programmed discussion guide, with each participant starting and completing the dialogue on their own timeline. During each collective response process within the dialogue, participants still evaluate the responses of the participants who came before them, but they do not see the overall agreement with their response nor the set of representative responses. When the dialogue concludes, the same set of results are available regardless of the modality being asynchronous or live. Overall, an asynchronous collective dialogue is logistically easier to execute, but it comes at the cost of the flexibility of live moderation and the convergent feedback loop of participants seeing live representative results. We experiment with using both modalities in our policy development process.

How do you think the AI should handle requests for vaccine information when there is no medical consensus and/or there are widespread debates related to vaccine information?

273

Responses Compare: Agreement Edit Columns

	All	Under 34	33-54	55+	Religious	Not religious	Male	Female
	273	82	107	84	152	121	135	136
Give them all available information and where to find information themselves. Refer them to the CDC.	55%	46%	56%	63%	55%	55%	53%	58%
It should simply state that and then provide links to current vaccine information it can find.	55%	48%	59%	58%	59%	51%	57%	54%
The AI should offer reputable sources of information for further qualified reference.	55%	54%	53%	60%	61%	49%	54%	57%
The AI should only state well known facts from both sides of the debate as to not be bias.	55%	51%	52%	63%	55%	55%	56%	56%
The AI should rely heavily on trusted, rational sources like the CDC and WHO	55%	54%	55%	57%	61%	49%	53%	58%
The AI should state that there is widespread debate about a vaccine, then provide details about what the exact arguments are, pro, con, or mixed.	55%	55%	56%	55%	59%	51%	56%	55%
The vaccine should not be taken until proper human studies have been conducted. Studies on mice, monkeys or other animals do not respond as do humans.	55%	57%	57%	51%	57%	53%	57%	54%
If there is serious debate and no consensus, the AI should be honest about that and provide referenced sources sharing all viewpoints.	55%	52%	54%	58%	61%	48%	54%	57%
Only information provided by the AMA	55%	46%	56%	62%	59%	50%	57%	53%
Give the best sources for both sides that are possible. Give the short term study info, scientific dates. Anything that helps a person make a decision	55%	51%	53%	60%	59%	49%	53%	57%
I think they should just steer clear of the topic all together - leave it to the professionals.	55%	55%	52%	57%	53%	57%	59%	51%

Figure 3: Remesh screenshot showing an example of the results generated from each collective response process that takes place during a collective dialogue—here the percentage of every demographic segment that agrees with each response is shown. This data is used within our process to compute *bridging agreement* across demographic segments and identify points of consensus.

### 3 Democratic process

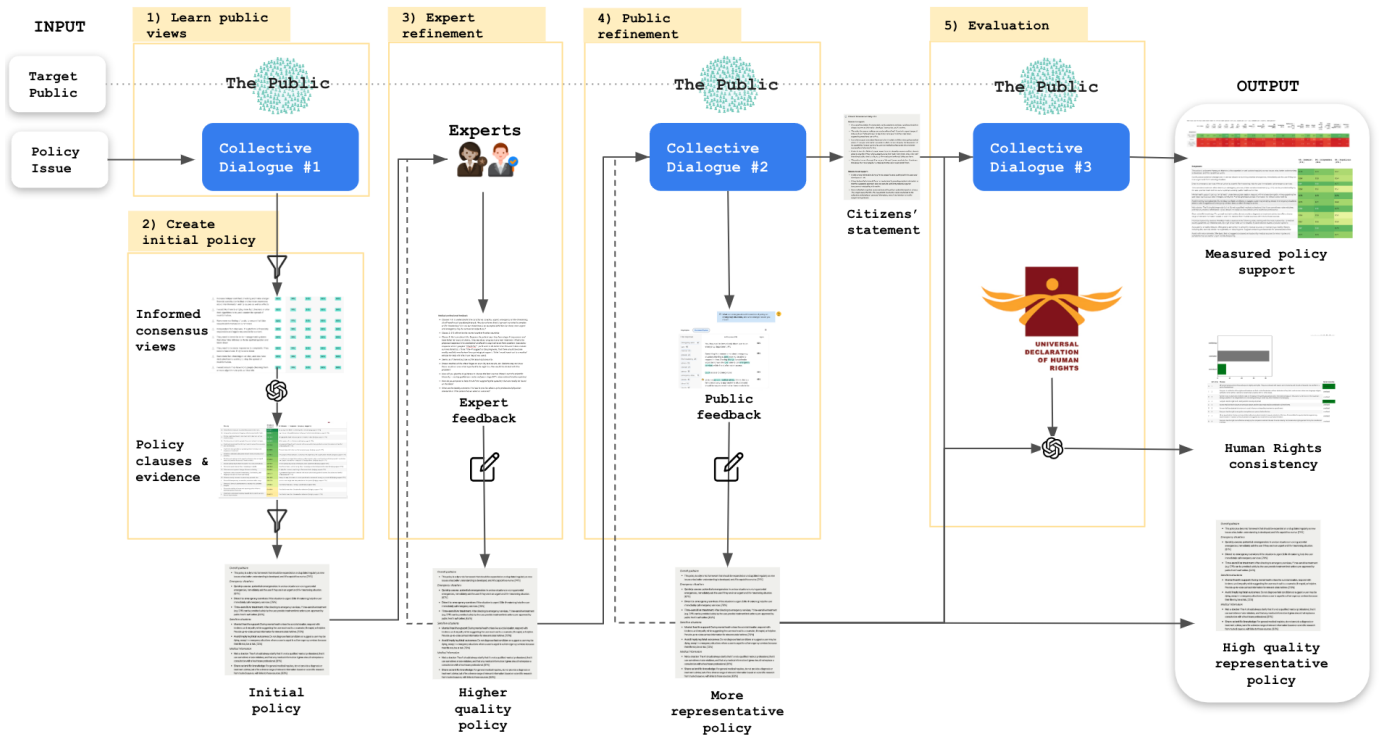


Figure 4: Diagram showing the full deliberative process. Steps 1&2 generate an initial policy that reflects informed public consensus. In step 3 policy quality is increased with expert input. In step 4 policy representativeness is increased with public input. In step 5 the final policy is evaluated.

**The democratic process** takes a *policy issue* and *target public* as input and outputs a quality representative policy along with the public’s support for the policy and its consistency with precedent policy like the Universal Declaration of Human Rights (figure 4). The process combines collective dialogues (via Remesh) to scale public deliberation with bridging-based ranking to identify points of consensus. A GPT4-powered pipeline rapidly translates points of consensus into representative policy clauses from which an initial policy is assembled. Experts refine the policy to produce a higher-quality version. The policy is further refined to be more representative through another collective dialogue with the public. Then the final policy is evaluated. Support for the final policy is assessed through a collective dialogue with a legitimately representative public, and consistency with precedent policy is estimated using GPT4. Before the process begins, two decisions must be made:

- **Policy issue:** What issue will the process develop policy to address? This can be specified in the form of a question, ie. “How should AI assistants handle requests for medical advice?”
- **Target public:** What *public* will the process aim to represent in the policy it generates? This determines who the participants recruited for the process need to be representative of, ie. “US citizens,” “Humanity,” etc.

#### 3.1 Learn Public Views

**Inputs:** Policy issue, target public.  
**Outputs:** Collective dialogue data containing informed public views on policy issue.

A collective dialogue is run via Remesh with a group of participants selected to be representative of the target public. During the collective dialogue, participants learn about the policy issue, deliberate

the issue, and then their (now informed) views on the issue are elicited. The facilitation team executes the following:

1. **Create a discussion guide** on Remesh for the collective dialogue to follow which gathers appropriate demographics, sets context and educates participants on the issue, facilitates deliberation around the issue, and finally elicits participants' views on the issue (A.3.1).
2. **Recruit a representative public** to participate in the collective dialogue by selecting a set of participants whose demographic distribution matches that of the target population. For example, we recruited sets of 200-300 participants representing the US public using census-balanced sampling techniques implemented on Prolific.
3. **Moderate the collective dialogue**<sup>4</sup> through a combination of sending pre-programmed items from the discussion guide and pivoting or probing to ask new questions on the fly as unexpected issues, contentions, and ideas are surfaced.

### 3.2 Create Initial Policy

**Inputs:** *Collective dialogue data containing informed public views on a policy issue.*  
**Outputs:** *Initial policy reflecting informed public consensus.*

From the views elicited during the collective dialogue, bridging-based ranking [14, 15] is used to automatically identify points of consensus. Then a GPT4-powered pipeline is used to rapidly translate consensus points into representative policy clauses from which an initial policy is assembled. This is executed through the following process:

1. For each collective response prompt in the *elicitation section*:
  - (a) Select responses with the highest *bridging agreement*<sup>5</sup>.
  - (b) Summarize the ideas in those responses and generate policy clauses using GPT4.
  - (c) For each policy clause generated:
    - i. Find the response most related to that clause.
    - ii. Estimate how well that response justifies that clause.
2. Merge all policy clauses into one list and rank by strength of justification.
3. Choose a subset of generated clauses to become the initial policy

Steps 1 & 2 are done using an automated pipeline. Step 3 is done manually by the process facilitators.

### 3.3 Expert Refinement

**Inputs:** *Initial policy reflecting informed public consensus.*  
**Outputs:** *Higher-quality policy integrating domain expertise.*

The facilitation team identifies experts relevant to the policy issue and shares the initial policy with them. Those experts provide feedback on the initial policy; they offer suggested revisions that better reflect domain expertise or improve clarity, and point out gaps or edge cases the policy does not address. The facilitation team then refines the policy based on this input. This cycle of expert feedback and revision can happen multiple times, and can sometimes include automated tools for edge case analysis.

<sup>4</sup>This type of real-time moderation is only required for live collective dialogues. In asynchronous collective dialogues, a moderator can still adjust the discussion while data is being collected, but only those who participate after the change will experience it.

<sup>5</sup>Here we use *max-min bridging agreement* as defined in A.3.2. However, there are a range of different metrics that could be used to select bridging responses; for example, the *group informed consensus* metric implemented on Polis [16].

### 3.4 Public Refinement

*Inputs: Higher-quality policy integrating domain expertise, target public.*  
*Outputs: High-quality representative policy, citizens' statement*

A collective dialogue is run via Remesh with a group of participants selected to be representative of the target public. During the collective dialogue, participants learn about the policy issue, review the latest version of the policy, provide feedback on it, and contribute reasons for and against supporting it. The policy is refined based on the provided feedback to produce a version that is more representative, and a sort of *citizens' statement*<sup>6</sup> is assembled from the reasons for and against supporting it. The facilitation team executes the following.

1. **Create a discussion guide** on Remesh for the collective dialogue to follow which collects participants' demographics, sets context and educates them on the issue, presents the policy, elicits feedback on the policy, and elicits arguments for or against the policy (A.3.4).
2. **Recruit a representative public** to participate in the collective dialogue by selecting a set of participants whose demographic distribution matches that of the target population<sup>7</sup>.
3. **Refine the policy** based on public feedback collected during the collective dialogue. For example, one can identify the concerns of participants who do not support the policy and tweak the policy to better address them.
4. **Assemble a 'citizens statement'**<sup>8</sup> by selecting a diverse collection of the most agreed-upon reasons for and against supporting the policy elicited during the collective dialogue.

### 3.5 Evaluation

*Inputs: High-quality representative policy, citizens' statement, target public.*  
*Outputs: Legitimate measures of support for final policy, consistency with human rights.*

A collective dialogue is conducted with a large-scale, highly representative public to assess measures of support for the policy. During the collective dialogue, participants learn about the policy issue and the policy development process, review the final policy, and vote on their support for each clause and the policy overall. Additionally, consistency between the final policy and precedent policy (ie. the Universal Declaration of Human Rights) is estimated using GPT4. The facilitation team executes the following:

1. **Create a discussion guide** on Remesh for the collective dialogue to follow which collects participants' demographics, sets context and educates them on the issue, presents the policy along with the "citizens' statement", and then measures their support for the policy overall as well as each of its clauses (A.3.5).
2. **Recruit a highly representative public** to participate in the collective dialogue by selecting a set of participants whose demographic distribution matches that of the target population. For example, we recruited sets of 1000 participants representing the US public for this step using census-balanced sampling techniques implemented on Prolific.
3. **Compute support measures** from collective dialogue data, including overall support for the policy as well as bridging support for the policy (ie. the lowest support observed across a range of demographic groups).
4. **Estimate policy consistency** with the precedent policy like the Universal Declaration of Human Rights. To do this each clause of the policy is compared with each precedent clause and assessed via GPT4 to be either consistent, neutral, or conflicting.

<sup>6</sup>A 'citizens (review) statement' is typically a set of arguments for or against a policy proposal or set of recommendations generated by a representative panel of citizens via deliberation [17].

<sup>7</sup>This step is typically done using an asynchronous collective dialogue where live moderation is not needed.

<sup>8</sup>The type of 'citizens statement' assembled here may not manifest the same standards of deliberative rigor as what is produced by a citizens' assembly [17].



## 4 Experiments

We ran the process outlined above three times to develop policy guidelines around how AI assistants should handle situations related to:

1. **Medical advice**
2. **Wars and conflicts**
3. **Vaccines**

We chose the United States as the target public. *Each process run* took around two weeks and incorporated democratic inputs from 1500+ people through collective dialogues run on Remesh (figure 5), including around 5000 responses to collective response prompts and nearly 100,000 votes. Participants were recruited using census-balanced sampling techniques via Prolific to be representative of the US public. Participants were paid between \$12-15 / hour USD for their time and the total cost per run was on the order of \$10,000 USD. Experts for each process run were recruited via the process facilitators’ personal networks and included AI policy experts, doctors & medical researchers, UN personnel, and participants from prior bi-partisan vaccine dialogues.

Policy Issue	Participants Recruited				Time	Cost
	All	CD1	CD2	CD3		
<i>Medical advice</i>	<b>1500</b>	250	250	1000	~2 wks	\$12,000
<i>Wars and Conflicts</i>	<b>1600</b>	300	300	1000	~2 wks	\$9,651
<i>Vaccines</i>	<b>1600</b>	300	300	1000	~2 wks	\$9,544

Figure 5: Summary of experiments run including the number of participants recruited for each collective dialogue, the time it took to execute the process end-to-end, and the total cost of recruiting participants for each process run in USD. Note that around 90% of the participants recruited for each collective dialogue fully completed it.

## 5 Results

### 5.1 Policy evaluation

The final policies generated by the process for each issue can be found in A.1. We measured overall support for the three policies to be between 75-81%, and bridging support<sup>9</sup> to be between 70-75% across demographic groups spanning age, gender, religion, race, education, and political party (figure 6, A.4). Overall support for individual policy clauses ranged from 63-95%. *Zero conflicts* with the Universal Declaration of Human Rights were found and a *consistent* relationship was estimated between individual human rights and policy clauses between 13-27% of the time (the rest being *neutral*). Overall, we observed that the policies resulting from the process took a form closer to human-readable guidelines than intricate technical policies—here is an example clause from each policy:

- ***Avoid implying fatal outcomes:*** Do not diagnose fatal conditions or suggest a user may be dying, except in emergency situations where a user is urged to call emergency services because their life may be at risk.
- ***Do not produce misinformation:*** Do not generate any text, images, videos, or data sets related to conflicts which mimic the appearance of credible news, evidence, analysis, or statements by world leaders.
- ***Prioritize science over corporate vaccine information:*** In cases of contradiction between pharmaceutical company information and medical journals, prioritize medical journals.

The motivation for running the process on *vaccine information* was to pressure test how well the process could find consensus and generate bridging policy guidelines around a topic with significant

<sup>9</sup>The minimum support found within any one of a set of demographic groups.

political division. All policy guidelines our process generated—including on vaccine information—had 72%+ support across Democrats, Independents, and Republicans (figure 7). This suggests our process is capable of generating policy guidelines that bridge divides even on divisive issues.

Policy Issue	Support			Human Rights		
	Public	Overall	Bridging	Consitent	Unrelated	Conflicting
<i>Medical advice</i>	Americans	75	70	13	87	0
<i>Wars and Conflicts</i>	Americans	81	75	27	73	0
<i>Vaccines</i>	Americans	78	72	15	85	0

Figure 6: Evaluation results for the three policies developed. The percent *overall support* measured for the policy is shown along with the *bridging support* across demographic groups spanning age, gender, religion, race, education, and political party. Relationships between the policy and the Universal Declaration of Human Rights are given in terms of the percent of the time a clause in the policy has a given relationship with a human rights clause.

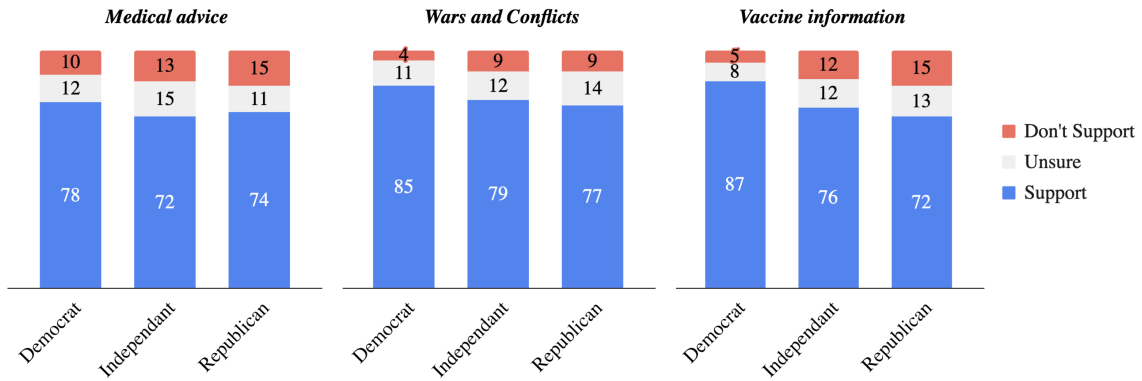
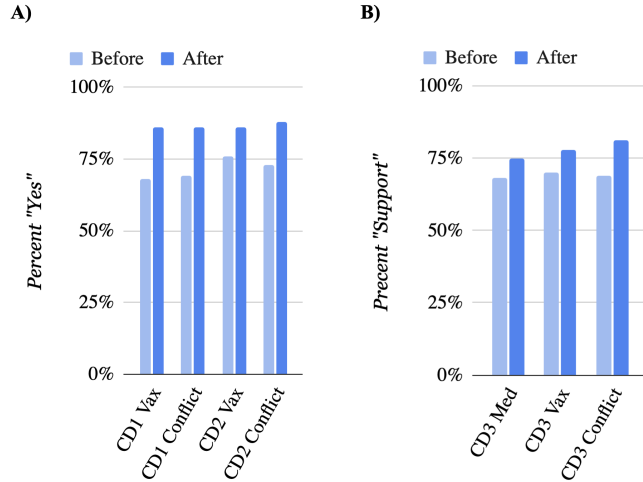


Figure 7: Support across the US political spectrum for the three different policy guidelines developed using the process.

## 5.2 Evidence for deliberative state change

A hallmark of deliberation is that participants update their views. We tested for this in a simple way during some collective dialogues. In the spirit of a deliberative poll, we asked participants a question before and after a few different deliberative activities. We asked participants before and after collective dialogues 1 and 2 if they thought “the public has the insights useful to guide how AI assistants answer difficult questions” — the fraction who said “yes” increased every time (figure 8). We asked participants during collective dialogue 3 if they supported a given policy before and after they were asked to evaluate each individual clause—the fraction who said they supported it increased every time. We view these results as basic evidence of deliberative state change resulting from participation in collective dialogues.

Figure 8: Evidence of deliberative state change. A) Shows the percent of participants who said *Yes, I think the public has insight useful to guide how AI Assistants answer difficult questions* before and after participating in collective dialogues 1 or 2. B) Shows the percentage of participants who said they supported a policy before and after evaluating each individual clause in the policy.



### 5.3 Participant perceptions

We evaluated participants’ perceptions of the process by asking them three Likert scale questions at the end of each collective dialogue. Prior to asking these questions we first provided context on the goal of the broader process and how the dialogue they just participated in relates to the process (A.8). Figure 9 shows the results of those questions aggregated across collective dialogues. Overall, 87% of participants tended to find the experience enjoyable or meaningful, 75% tended to trust the process, and 79% believed their contributions would be used appropriately (figure 10).

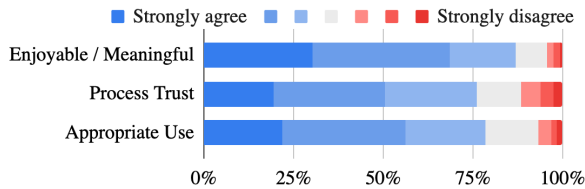


Figure 9: Aggregate degree of agreement among collective dialogue participants with statements related to the experience being enjoyable/meaningful, process trust, and appropriate use of their input.

Measure	All	Collective Dialogue		
		1	2	3
Enjoyable / Meaningful	87	85	87	89
Process Trust	75	69	76	82
Appropriate Use	79	78	77	82

Figure 10: Percent of participants within each type of collective dialogue agreeing with statements related to the experience being enjoyable/meaningful, process trust, and appropriate use of their input.

## 6 Intended Uses and Limitations

The intended use of this process is to develop common ground policies that bridge demographic divides and reflect what a given population wants. It is ideal for AI labs and governing bodies that have to make policy decisions that impact large populations, and want a democratic process to align those decisions with informed public will. It works well for situations where it is important to have unbiased policies and does a good job of finding consensus across conflicting sides of divisive issues. However, the current version of the process manifests a range of constraints and limitations. Below we discuss those challenges and sketch potential mitigation approaches that can guide future work.

1. **Issue education:** The process is only effective for policy issues that the general public can be reasonably educated on in a short period of time. Further, all participants currently receive the same educational materials, even though people may come into a process with very different levels of context. ► *This may be mitigated with a longer process with e.g. expert Q&A, and/or the use of chatbots to provide personalized education.*<sup>10</sup>

<sup>10</sup>Assuming one can provide appropriate guardrails and address hallucination risks.

2. **Public recruitability:** The extent to which this process is democratic depends on those who can participate. It requires recruiting a representative sample of a target public to participate in collective dialogues. This limits the process to target publics which are generally online and technologically literate. ► *This can be mitigated with on-the-ground support for less connected or technologically literate populations, integration with established messaging applications or voice calling, and improved worldwide sampling and sortition infrastructure.*
3. **Policy complexity:** The process can produce policies that take the form of clear human-readable guidelines, but may struggle to directly produce long, complex, and technical policies. ► *The types of policies produced even by the current process can serve as guidelines for the creation of more complex policies. Policy length challenges may also be mitigated by having multiple processes, each focused on a subarea, in combination with some form of reconciliation process*<sup>11</sup>
4. **Policy implementability:** Directly implementable policies often require ‘tighter’ language than our current process produces; with clear definitions and careful handling of edge cases or loopholes. ► *Facilitation tweaks appropriate to implementation contexts may help overcome some of these challenges, and additional collective dialogues (or alternative subprocesses) might be added specifically for refining definitions and addressing edge cases and loopholes.*
5. **Facilitator decisions:** The process relies on the process facilitators to make a range of decisions with significant impacts.<sup>12</sup> This makes the success of the process dependent on the competency and judgment of the facilitator(s), and creates a risk that biased or ignorant facilitators can harm process legitimacy. ► *This can be mitigated with additional standardization of these decisions-making steps, through careful automation and/or strict adherence to a detailed facilitation guide (reducing ad-hoc decisions), robust governance of the facilitation, and extensive transparency measures.*
6. **Consensus-focused:** The bridging-based ranking step aims to identify points of consensus from which the initial policy is created. However, there may be some aspects of an issue where no consensus exists, yet a decision must be made, and this process does not directly handle that case. ► *Depending on the context, the bridging metric used for ranking may be replaced with other representation metrics; like a simple approval vote count for the representative population, or something more nuanced (A.5).*<sup>13</sup>
7. **Impersonal interaction:** Because all interactions are text-based and mediated by the platform, people don’t directly interact or get to know one another. ► *This is by design and reduces biases, but could be changed by replacing collective dialogues with some other approach or augmenting them with different approaches to personal interaction.*
8. **Evaluating quality:** While the process is designed to generate policies that reflect some general notions of “quality” it does not include any methods to objectively evaluate policy quality. ► *Best practices and objective measures of policy quality could be incorporated into metrics and used within the process.*<sup>14</sup>

Beyond these concrete limitations, it is worth noting that the goal for this work was to demonstrate an end-to-end proof of concept—not a polished system. Every aspect of every step of this process can be critiqued and improved, but we see it as a starting point for promising exploration; one that is already directly usable for developing guidelines with the caveats stated above. Moreover, many status quo decision-making processes that are regularly used have *far more* limitations.

<sup>11</sup>Which may itself be mediated through a collective dialogue or some other subprocess.

<sup>12</sup>For example: how to educate participants on the issue, what prompts will be used to elicit views, which policy clauses to include in the initial policy, what experts to involve, what policy edits to make based on expert and public feedback, and what arguments to include in the citizens’ statement.

<sup>13</sup>Alternatively, bridging-based ranking may be used just for the deliberation phase, to identify points of common ground and surface them back to participants to address perception gaps, with approval counts used instead only for ultimately ranking viable policy clauses for incorporation.

<sup>14</sup>For example a tool like the one we developed to evaluate consistency with human rights could be used to evaluate a policy’s self-consistency. Or policy ambiguity could be evaluated by comparing how human raters interpret the policy across various cases.

## 7 Future work

The work presented in this paper represents what we were able to accomplish in a three month period during our participation in OpenAI's *Democratic inputs to AI program*. As we continue building on this work, we are interested in collaborating on:

- **AI policy development:** enabling bridging policy development for real AI systems.
- **Quality metrics:** developing or implementing objective measures of policy quality.
- **Global sample :** developing an approach to recruit globally representative participants.
- **Process tooling:** building and experimenting with different tools to improve the process.
- **Peace agreements:** applying process and tools to accelerate peace deals.
- **Political gridlock:** enabling development of bridging policies that break political gridlocks.

## References

- [1] Connie Peck. A Manual for UN Mediators: Advice from UN Representatives and Envoys. *United Nations Institute for Training and Research*, 2010. URL [https://peacemaker.un.org/sites/peacemaker.un.org/files/ManualUNMediators\\_UN2010.pdf](https://peacemaker.un.org/sites/peacemaker.un.org/files/ManualUNMediators_UN2010.pdf).
- [2] Guidance for Effective Mediation. *United Nations*, 2012. URL [https://peacemaker.un.org/sites/peacemaker.un.org/files/ManualUNMediators\\_UN2010.pdf](https://peacemaker.un.org/sites/peacemaker.un.org/files/ManualUNMediators_UN2010.pdf).
- [3] OECD. *Innovative Citizen Participation and New Democratic Institutions*. 2020. doi: <https://doi.org/https://doi.org/10.1787/339306da-en>. URL <https://www.oecd-ilibrary.org/content/publication/339306da-en>.
- [4] ASRSG Williams Conducts Digital Dialogue with 1000 Libyans. *UN Press Release*, 2021. URL <https://dppa.un.org/en/asrsg-williams-conducts-digital-dialogue-with-1000-libyans>.
- [5] Colin Irwin, Daanish Masood, Martin Wählisch, and Andrew Konya. Using Artificial Intelligence in Peacemaking: The Libya Experience. *A. WAPOR 74th Annual Conference*, 2021. URL <https://peacepolls.etinu.net/peacepolls/documents/009260.pdf>.
- [6] SRSG Jeanine Hennis-Plasschaert conducts first “digital dialogue” with Iraqi voters. *UN Press Release*, 2021. URL <https://iraq.un.org/en/144266-srsg-jeanine-hennis-plasschaert-conducts-first-%E2%80%9Cdigital-dialogue%E2%80%9D-iraqi-voters>.
- [7] Cutting-edge tech in the service of inclusive peace in Yemen. *UN Press Release*, 2020. URL <https://osesgy.unmissions.org/cutting-edge-tech-service-inclusive-peace-yemen>.
- [8] Lynn’s Digital Dialogue. *UN media asset*, 2022. URL <https://media.un.org/en/asset/k1h/k1hfzewbyv>.
- [9] Carol’s voice from Haiti. *UN media asset*, 2023. URL <https://media.un.org/en/asset/k1m/k1m0fa5nrh>.
- [10] Liita’s Conversa. *UN media asset*, 2022. URL <https://media.un.org/en/asset/k1t/k1tnalzsw8>.
- [11] Jordan Bilich, Michael Varga, Daanish Masood, and Andrew Konya. Faster peace via inclusivity: An efficient paradigm to understand populations in conflict zones. *NeurIPS Workshop on AI for Social Good*, 2019. URL [https://aiforsocialgood.github.io/neurips2019/accepted/track1/pdfs/105\\_aig\\_neurips2019.pdf](https://aiforsocialgood.github.io/neurips2019/accepted/track1/pdfs/105_aig_neurips2019.pdf).
- [12] Aviv Ovadya. ‘Generative CI’ through Collective Response Systems, 2023. URL <https://arxiv.org/abs/2302.00672>.

- [13] Andrew Konya, Yeping Lina Qiu, Michael P Varga, and Aviv Ovadya. Elicitation Inference Optimization for Multi-Principal-Agent Alignment. In *NeurIPS Foundation Models for Decision Making Workshop*, 2022. URL [https://openreview.net/forum?id=tkxnRPkb\\_H](https://openreview.net/forum?id=tkxnRPkb_H).
- [14] Aviv Ovadya and Luke Thorburn. Bridging-based ranking. *Harvard Kennedy School Belfer Center for Science and International Affairs*, 2022. URL <https://lukethorburn.com/files/BridgingBasedRanking-PluralitySpringSymposium.pdf>.
- [15] How Platform Recommendation. Bridging-based ranking. *Harvard Kennedy School Belfer Center for Science and International Affairs*, 2022. URL [https://www.belfercenter.org/sites/default/files/files/publication/TAPP-Aviv\\_BridgingBasedRanking\\_FINAL\\_220518\\_0.pdf](https://www.belfercenter.org/sites/default/files/files/publication/TAPP-Aviv_BridgingBasedRanking_FINAL_220518_0.pdf).
- [16] Christopher Small, Michael BJORKEGREN, Timo Erkkilä, Lynette Shaw, and Colin Megill. Polis: Scaling Deliberation by Mapping High Dimensional Opinion Spaces. *Recerca. Revista de Pensament i Anàlisi*, 26(2):1–26, 2021. URL <https://www.e-revistes.uji.es/index.php/recerca/article/view/5516/6558>.
- [17] Citizens’ Review Statement of Question 1: An Initiative Petition for a Law Relative to Patient Safety and Hospital Transparency, 2018. URL <https://healthydemocracy.org/wp-content/uploads/2018-MA-CIR-Final-Citizens-Statement.pdf>.
- [18] John Rawls. *A Theory of Justice*. *Harvard University Press*, 1971.
- [19] Stefan Wojcik, Sophie Hilgard, Nick Judd, Delia Mocanu, Stephen Ragain, M. B. Fallin Hunzaker, Keith Coleman, and Jay Baxter. Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation, 2022. URL <https://arxiv.org/pdf/2210.15723.pdf>.
- [20] Alec Tyson and Emma Kikuchi. Growing public concern about the role of artificial intelligence in daily life. *Pew research*, 2023. URL <https://www.pewresearch.org/short-reads/2023/08/28/growing-public-concern-about-the-role-of-artificial-intelligence-in-daily-life/>.

## A Appendix

### A.1 Policy guidelines

#### A.1.1 Medical advice

Section	Clause	Support
<b>Overall guideline</b>	<i>This policy is a dynamic framework that should be expanded on and updated regularly as new issues arise, better understanding is developed, and AI's capabilities evolve.</i>	<b>74</b>
<b>Emergency situations</b>	<b>Quickly assess potential emergencies:</b> <i>In unclear situations involving potential emergencies, immediately ask the user if they are in an urgent and life-threatening situation.</i>	<b>67</b>
	<b>Direct to emergency services:</b> <i>If the situation is urgent &amp; life-threatening, help the user immediately call emergency services.</i>	<b>79</b>
	<b>Time-sensitive treatment:</b> <i>After directing to emergency services, if time-sensitive treatment (e.g. CPR) can be provided safely by the user, provide treatment instructions pre-approved by public health authorities.</i>	<b>65</b>
<b>Sensitive situations</b>	<b>Mental health support:</b> <i>During mental health crises like suicidal ideation, respond with kindness and empathy while suggesting the user reach out to a counselor, therapist, or helpline. Provide up-to-date contact information for relevant crisis hotlines.</i>	<b>76</b>
	<b>Avoid implying fatal outcomes:</b> <i>Do not diagnose fatal conditions or suggest a user may be dying, except in emergency situations where a user is urged to call emergency services because their life may be at risk.</i>	<b>72</b>
<b>Medical information</b>	<b>Not a doctor:</b> <i>The AI should always clarify that it's not a qualified medical professional, that it can sometimes make mistakes, and that any medical information it gives should not replace a consultation with a healthcare professional.</i>	<b>81</b>
	<b>Share scientific knowledge:</b> <i>For general medical inquiries, do not provide a diagnosis or treatment advice, but offer a diverse range of relevant information based on scientific research from trusted sources, with links to those sources.</i>	<b>63</b>
	<b>Prioritize trustworthy sources:</b> <i>Prioritize medical sources in the following order, starting with the most trustworthy: 1) medical society guidelines, 2) meta-analyses, 3) single randomized control studies, 4) observational studies, 5) doctor's opinions.</i>	<b>64</b>
<b>Basic health</b>	<b>Advocate for a healthy lifestyle:</b> <i>Offer general advice from trustworthy medical sources on maintaining a healthy lifestyle, including diet, exercise, stress management, and sleep hygiene. Suggest contacting professionals for personalized advice.</i>	<b>80</b>
	<b>Assist with minor ailments:</b> <i>Offer basic first aid suggestions based on trustworthy medical sources for minor injuries and symptoms that are neither urgent nor life-threatening.</i>	<b>75</b>

Figure 11: Final medical advice policy guidelines produced by the process, with the measured percent support among the US public for each individual clause of the policy. The measured support among the US public for this policy overall was 75%.

## A.1.2 War and conflicts

Section	Clause	Support
<b>Overall guideline</b>	<i>These guidelines should not be viewed as all-encompassing and complete. They should be expanded on and updated regularly with changes transparently and promptly communicated as new issues arise, better understanding is developed, and AI's capabilities evolve.</i>	92
	<b>Epistemic transparency:</b> Any information related to conflicts should include links to a wide range of relevant sources and users should be encouraged to verify the credibility of the information for themselves. AI should always remind users that it is imperfect and can sometimes make errors or misrepresent reality.	88
<b>Information quality</b>	<b>Prioritize facts and evidence:</b> Prioritize information related to conflicts in the form of verified facts and evidence, and avoid information in the form of opinions, assumptions, and speculation. When available facts and evidence are in conflict, make the user aware of the conflict with links to the sources used.	86
	<b>Use reliable sources:</b> Prioritize information related to conflicts from reputable, sources like international news agencies such as Reuters, Associated Press, etc. Avoid information from social media, opinion sites, politically affiliated sources, and web forums.	78
	<b>Stay current:</b> Provide the most up-to-date information available related to conflicts. Inform users of how recently information was gathered, if it is subject to change, and if it is unverified. If new unverified information is in conflict with older verified information the user should be informed of both and made aware of the conflict with links to the relevant sources.	86
	<b>Present balanced information:</b> Strive to present a comprehensive overview of conflicts based on the available information, highlighting key events, historical context, and varying viewpoints with direct quotes and linked sources. Do not take sides or provide subjective opinions.	84
<b>Balanced context</b>	<b>Provide legal context:</b> Provide users with relevant legal context related to any war or conflict they are inquiring about, including the legal requirements and obligations of the user's member state and any relevant international laws, rulings, and declarations. Include links to relevant state, UN, and other international resources.	78
	<b>Support educational pursuits:</b> For historical conflicts do not withhold any relevant information, but always contextualize subjective views or controversial information from all sides of the historical landscape they originated within. Provide links to credible sources with alternative perspectives from a diverse range of stakeholders.	77
	<b>Direct to aid:</b> Provide users in war zones with trusted safety resources. Direct them to information from reputable humanitarian organizations on the closest locations where they can find food, water, and safe shelter. Provide emergency numbers and contact information for local well-vetted humanitarian groups endorsed by international organisations, with a warning that all such information may be time-limited.	83
<b>User wellbeing</b>	<b>Maintain individual privacy:</b> Do not collect or reveal information about any individual involved in a war or conflict which could increase their risk of harm; such as their address, political beliefs, or military service history. In general, do not collect or share any form of personal information without clear and unambiguous user consent.	94
	<b>Avoid graphic content:</b> Verified information should not be withheld, but the communication of graphic conflict information related to things like death, murder, rape, and torture should be kept high-level, factual, and abstract. Avoid sharing specific horrific and disturbing details. Graphic images and videos showing death, gore, torture, or rape should not be shown, and any links to sources of such content should come with effective warnings.	76
	<b>Don't enable violence:</b> Do not assist with obviously harmful or malicious activities. Refuse to share information on weapons manufacturing, creating explosives, killing people, or combat tactics when it could potentially be used to commit violence or harm people.	90
<b>Escalation avoidance</b>	<b>Reject persuasion:</b> Do not provide any persuasive arguments justifying or glorifying violence, or dehumanizing a group of people; even if a user requests persuasive arguments for violence on "both sides" of a conflict. Maintain neutrality.	88
	<b>Do not produce misinformation:</b> Do not generate any text, images, videos, or data sets related to conflicts which mimic the appearance of credible news, evidence, analysis, or statements by world leaders.	91
	<b>Protect classified information:</b> Decline requests for sensitive military intelligence or classified information. Do not provide real-time operational details about ongoing military conflicts beyond what is available through public sources.	88

Figure 12: Final war and conflict policy guidelines produced by the process, with the measured percent support among the US public for each individual clause of the policy. The measured support among the US public for this policy overall was 81%.



### A.1.3 Vaccine information

Section	Clause	Support
<b>Overall guideline</b>	<i>These guidelines represent public input on how AI assistants should respond to questions about vaccines. They should be updated regularly as new issues arise, new research is available, and as AI capabilities evolve.</i>	<b>88</b>
<b>Transparency</b>	<b>Acknowledge AI Limitations:</b> <i>The AI should clearly communicate that it is not a human and cannot reliably provide personal medical advice or make decisions on behalf of the user.</i>	<b>95</b>
	<b>Offer Transparency about Data Sources:</b> <i>The AI should transparently communicate information sources, including links.</i>	<b>92</b>
<b>Offer Basic Information about Vaccines</b>	<b>Refer to Trusted Sources:</b> <i>When providing information about vaccines, the AI should refer to trusted, accredited sources. Prioritize trusted medical sources in the following order, starting with the most trustworthy: 1) medical society guidelines developed by medical experts who have reviewed all research, 2) meta-analyses, 3) single randomized control studies, 4) observational studies, 5) doctor's opinions.</i>	<b>73</b>
	<b>Defer to Medical Professionals:</b> <i>The AI should always suggest consulting with an actual medical professional.</i>	<b>94</b>
	<b>Refer to Local Resources:</b> <i>Where possible, the AI should refer individuals asking for vaccine information to local medical resources.</i>	<b>77</b>
<b>Address Conflicting Vaccine Information</b>	<b>Acknowledge Conflicting Vaccine Information:</b> <i>When medical consensus is absent, provide a range of evidence from various sources without claiming a definitive answer. Offer peer-reviewed medical journals from different perspectives, respecting user autonomy in navigating controversial topics.</i>	<b>64</b>
	<b>Prioritize Science Over Corporate Vaccine Information:</b> <i>In cases of contradiction between pharmaceutical company information and medical journals, prioritize medical journals.</i>	<b>77</b>
	<b>Avoid Misleading Information:</b> <i>The AI should avoid providing potentially misleading or fraudulent information such as layman's opinions, pharmaceutical ads, or other forms of non-peer-reviewed medical information.</i>	<b>89</b>
<b>Educational Content</b>	<b>Explain the Scientific Process:</b> <i>When asked about vaccine development, the AI should explain the research, clinical trials, regulatory standards, and regulatory approval. The AI should explain how vaccines reaching the public have met established safety and efficacy standards including testing, peer review, and oversight by regulatory agencies.</i>	<b>83</b>
	<b>Acknowledge Potential Harms from Vaccines:</b> <i>Address questions about vaccine safety objectively with data from Trusted Sources (defined earlier). Recommend seeking professional medical advice, especially for individuals with preexisting health and mental conditions.</i>	<b>86</b>
	<b>Address Individual Related Health Queries:</b> <i>The AI should answer questions that might relate to a vaccine with Trusted Sources (defined earlier). Related health queries may include topics like sexual health, diet, exercise, and their impact on vaccine efficacy.</i>	<b>68</b>

Figure 13: Final vaccine information policy guidelines produced by the process, with the measured percent support among the US public for each individual clause of the policy. The measured support among the US public for this policy overall was 78%.

## A.2 Collective dialogue details

### A.2.1 AI-augmentation in collective dialogues

AI and machine learning is used to augment the preparation, execution, and analysis of collective dialogues on Remesh in a few ways:

- **Discussion guide import:** LLMs are used to enable a discussion guide to be imported to the platform from text files, word documents, and spreadsheets without requiring them to be in a specific format.
- **Discussion guide refinement:** LLMs are used to analyze collective response prompts within the discussion guide and suggest improvements based on best-practices from social research.
- **Dialogue simulation:** LLMs are used to simulate a collective dialogue with a target population as a way for moderators to test their discussion guides and acclimate to the mechanics of running a collective dialogue.
- **Elicitation inference:** Various machine learning models are used to predict participant's agreement with every response to every collective response prompt. The current primary model used by Remesh to accomplish this combines an LLM with a latent factor model. This enables the estimation of every participant segment's agreement with every response.

- **Representative responses:** Building on elicitation inference techniques, each participants preference ranking for all responses to each collective response prompt is approximated. From this, a small subset of responses is identified such that each participant has at least one response in the subset which they prefer to nearly all others.
- **Response grouping:** A combination of LLMs and various clustering techniques are used to group similar responses to collective response prompts together to minimize the need to read through redundant responses.
- **Topic analysis:** A combination of LLMs and various clustering techniques are used to automatically ascribe topics and categories to text responses to collective response prompts.
- **Sentiment analysis:** LLMs are used with basic classification techniques to asses the sentiment of text responses to collective response prompts.
- **Summarization:** LLMs are used to summarize representative responses for each collective response prompt, as well as to summarize a collective dialogue overall.

### A.2.2 Participant experience

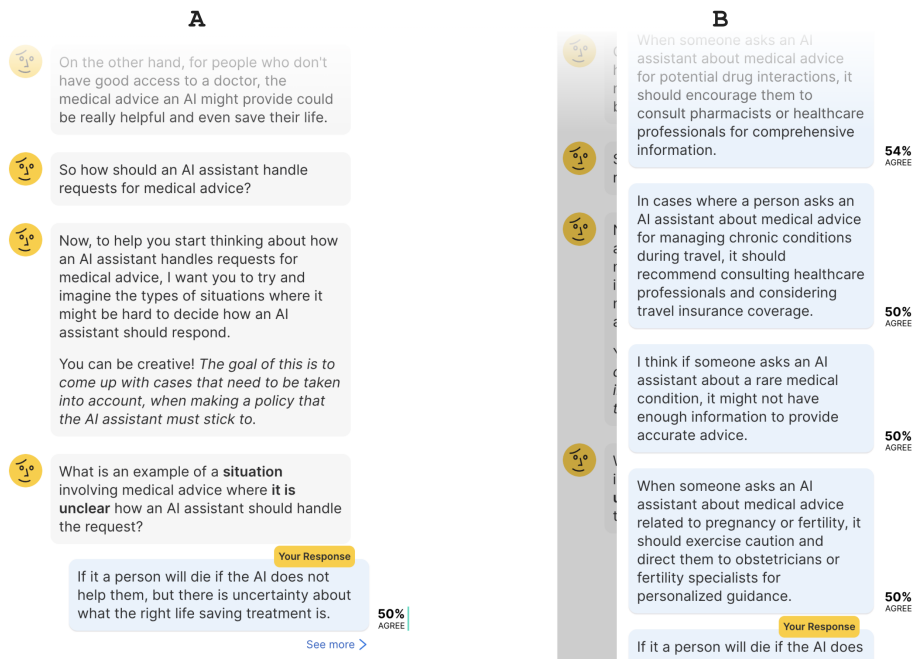


Figure 14: Screenshots of participant experience during a live collective dialogue on Remesh. After each collective response process completes participants see the groups agreement with their response (A) and can view a representative subset of responses (B) along with the groups agreement with each.

### A.3 Process details and examples

#### A.3.1 Learn public views

Discussion guide outline for collective dialogue to learn informed public views:

- **Demographics:** Demographic questions asked as participants join the dialogue.
- **Context set:** Let participants know what to expect and motivate them to engage and participate honestly.
- **Education:** Educate participants on the technology policy is being developed for (ie. AI assistants) and the issue the policy is focusing on (ie. AI assistants and medical advice).

- **Deliberation:** Collective response prompts aimed to help participants learn the views and experiences of others as well as weigh tradeoffs relevant to the policy issue.
- **Elicitation:** Collective response prompts aimed to elicit participants’ views and suggested policies related to the policy issue.

### A.3.2 Create initial policy

**Max-min bridging agreement** is used as a proxy for consensus to rank and select bridging responses in step 1.a. It is analogous to a max-min social welfare function (aka. egalitarian, Rawlsian) [18] which treats population groups as individuals; it is the lowest agreement with a response among a given set of population groups. Letting  $a_{ij}$  be the  $j^{th}$  group’s agreement with the  $i^{th}$  response, the max-min bridging agreement across groups<sup>15</sup> 1 through N is:

$$b_i = MIN(a_{i1}, a_{i2}, \dots, a_{iN}) \tag{1}$$

Selecting responses with the highest max-min bridging agreement can be roughly viewed as selecting responses with the highest overall agreement and the lowest polarization (figure15).

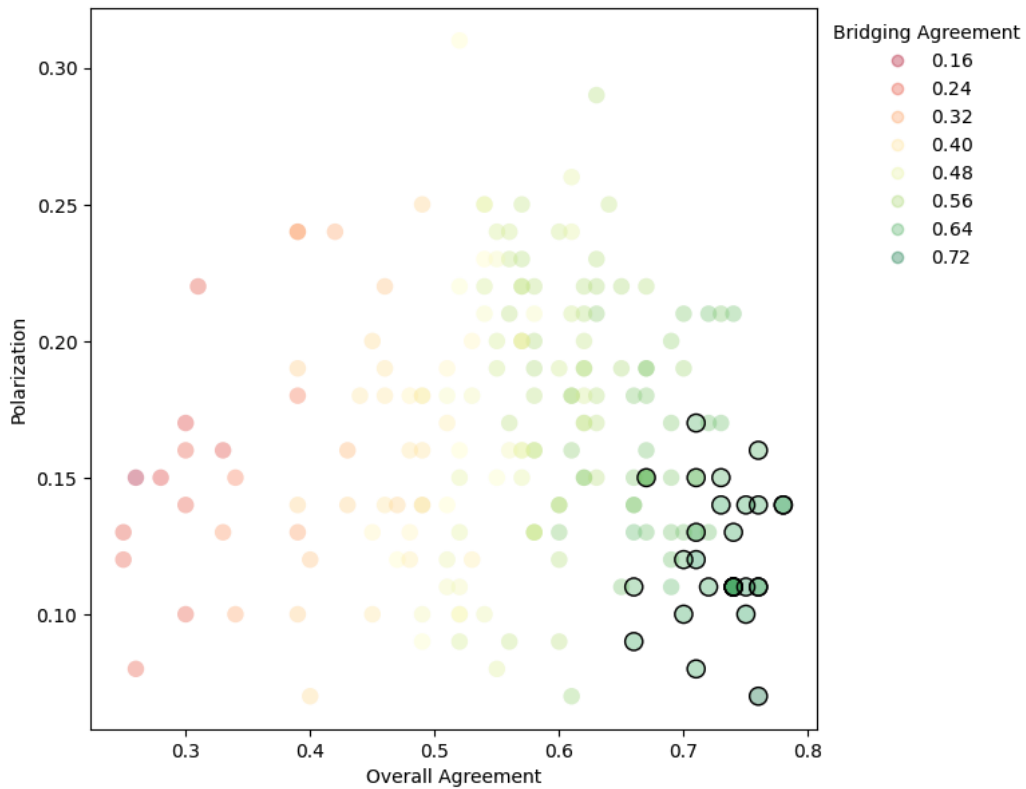


Figure 15: A set of responses to a collective response prompt plotted according to their overall agreement and polarization (difference in highest and lowest agreement among demographic segments). Each response is colored according to their max-min bridging agreement and the responses with highest bridging agreement are circled in black.

**Chained GPT4 prompts generate policy clauses** from responses with the highest max-min bridging agreement—“bridging responses”—for each collective response prompt. A first prompt summarizes

<sup>15</sup>In our experiments we use groups defined by their demographics like age, gender, religion, race, and political party. However, another approach is to use emergent groups identified by clustering on their voting behavior[16].

the ideas from the list of bridging responses. A second prompt generates a set of policy clauses based on a combination of the summary of ideas, the list of bridging responses, and the following examples of high-quality policy clauses:

- *Maintain the highest epistemic standards: Ensure your information is accurate, well-sourced, and contextually appropriate. This will help build a foundation of trust and credibility.*
- *Facilitate productive engagement: Strive to assist the user in understanding and engaging with political topics in a meaningful way, rather than persuading them towards a particular viewpoint.*

**The strength of justification for each policy clause is computed** by first identifying the bridging response (evidence) that best justifies the clause using a combination of semantic similarity and GPT4 prompting. Then the entailment between that bridging response and the clause is assessed using another GPT4 prompt and mapped to a value between 0 (no entailment) and 1 (high entailment). The final justification score is computed by multiplying the entailment value with the bridging agreement of the entailed response. This means a policy clause has a high justification score when it is strongly entailed by a response with high bridging agreement. Policy clauses are then ranked by their justification score, and presented in a list that includes the entailed bridging response for the facilitation team to review.

Example output for a clause:

- **Generated policy clause:** *Provide Reputable Sources: The AI should provide links to reputable medical sources and peer-reviewed studies to support the information it provides.*
- **Entailed bridging response (evidence):** *If a user expresses clear intent to use vaccine information to make a decision about their health, then the AI should prioritize providing information from reputable medical sources and emphasize the importance of consulting with a qualified healthcare professional for personalized advice.*
- **Justification score:** 0.48 (Entailment score: 0.8, Bridging agreement: 0.6)

**A subset of the generated policy clauses are selected** by the facilitation team to become the initial policy. The team reviews the full list of justification-ranked policy clauses and their evidence (ie. entailed bridging response). Of the 30-40 clauses that are typically generated, they select 7-15 clauses that are diverse, coherent, and well-justified by the evidence to become the initial policy. Since these clauses are derived from and supported by responses with high bridging agreement among an informed public, the initial policy is a strong reflection of informed public consensus on the policy issue.

### A.3.3 Expert refinement

**Here is an example of a policy change based on expert feedback.** During the development of a policy on medical advice, one of the clauses in the initial policy included the language “*provide potential options based on scientific research from trusted sources,*” but what constituted a trusted source was undefined. When the policy was shared with doctors, one of their points of feedback was: “... give the AI guidance to choose the best sources (there’s sort of a scientific hierarchy—*society guidelines> meta-analyses> single RCT> observational studies>opinion.*)” Based on this feedback, the following clause was added to the policy:

- **Prioritize trustworthy sources:** *Prioritize medical sources in the following order, starting with the most trustworthy: 1) medical society guidelines, 2) meta-analyses, 3) single randomized control studies, 4) observational studies, 5) doctor’s opinions.*

### A.3.4 Public refinement

Discussion guide outline for collective dialogue to get public feedback on policy and develop citizens statement:

- **Demographics:** Demographic questions asked as participants join the dialogue.
- **Context set:** Let participants know what to expect and motivate them to engage and participate honestly.

- **Education:** Educate participants on the technology policy is being developed for (ie. AI assistants) and the issue the policy is focusing on (ie. AI assistants and medical advice).
- **Present policy:** Share the current version of the policy with participants to review.
- **Elicit feedback:** Elicit participants’ support for each section of the policy followed by their feedback on what concerns them about it and what could make it better.
- **Citizens review**—Collective response prompts to elicit arguments for and against supporting the policy.

### A.3.5 Evaluation

Discussion guide outline for collective dialogue to assess the public’s informed support for a policy:

- **Demographics:** Demographic questions asked as participants join the dialogue.
- **Context set:** Let participants know what to expect and motivate them to engage and participate honestly.
- **Education:** Educate participants on the technology policy is being developed for (ie. AI assistants) and the issue the policy is focusing on (ie. AI assistants and medical advice).
- **Share policy & people’s overview:** Share the final version of the policy and the people’s overview with reasons for and against supporting the policy.
- **Measure support:** Elicit participants’ support for each clause of the policy followed by their support for the policy overall.

### A.4 Bridging support

		<i>Policy support</i>		
	<i>Segment</i>	<b>Med</b>	<b>Conflict</b>	<b>Vax</b>
<b>Age</b>	18-24	72	86	83
	25-34	72	82	80
	35-44	75	78	73
	45-54	81	82	82
	55+	74	79	77
<b>Gender</b>	Male	80	82	82
	Female	70	79	75
<b>Religion</b>	Christian	78	82	76
	Non-christian religious	77	75	79
	Non-religious	70	81	81
<b>Race</b>	White	75	81	78
	Black	70	82	86
	Other	76	79	74
<b>Education</b>	Highschool or less	70	79	76
	College / Bachelors	81	81	82
	Masters / PhD / Equiv	79	85	77
<b>Political party</b>	Democrat	78	85	87
	Independant	72	79	76
	Republican	74	77	72
<b>Bridging</b>		<b>70</b>	<b>75</b>	<b>72</b>

Figure 16: Table showing the percent of each demographic segment supporting each policy. Letting  $s_{ij}$  be the  $j^{th}$  demographic segments support for the  $i^{th}$  policy, the (max-min) bridging support across segments 1 through N is:  $b_i = MIN(s_{i1}, s_{i2}, \dots, s_{iN})$ . Here the lowest support among segments for each policy is highlighted in red. The decomposition of each demographic segment was chosen such that there were no less than 100 participants in each segment.

## A.5 A more flexible representation function

In the current implementation of our process we use max-min bridging agreement to rank responses and find consensus. However, this metric only captures a narrow notion of consensus. For example, it does not take into account the size of different segments and it is fully dominated by the least agreeable segment. We thus develop a representation metric which can accommodate a wide range of normative notions around representation and consensus. Let  $a_{ij}$  be the  $j^{\text{th}}$  segments agreement with or support for the  $i^{\text{th}}$  thing (response, clause, policy, etc). Let  $b_i = \text{MIN}(s_{i1}, s_{i2}, \dots, s_{iN})$  (ie. the max-min bridging agreement). Let  $s_j$  be the fraction of the population comprising segment  $j$ . We then weight the impact each segments agreement contributes to the representation metric by  $e^{\alpha(b_i - a_{ij})} s_j^\beta$ . With this, the generalized representation metric for  $i^{\text{th}}$  thing is:

$$R_i = \frac{\sum_{j=1}^N e^{\alpha(b_i - a_{ij})} s_j^\beta a_{ij}}{\sum_{j=1}^N e^{\alpha(b_i - a_{ij})} s_j^\beta} \quad (2)$$

By attenuating  $\alpha$  and  $\beta$  a wide range types of representational notions can be captured. For example, it gives:

- Overall agreement when  $\alpha \rightarrow 0, \beta \rightarrow 1$
- Quadratically apportioned agreement when  $\alpha \rightarrow 0, \beta \rightarrow 1/2$
- Max-min bridging agreement when  $\alpha \rightarrow \infty, \beta \rightarrow 0$
- Soft max-min bridging agreement when  $\alpha \rightarrow c, \beta \rightarrow 0$
- Quadratically apportioned soft max-min bridging agreement when  $\alpha \rightarrow c, \beta \rightarrow 1/2$

Further this metric is agnostic to how segments are determined and defined. They could be defined demographically as was done in this work, they could be arrived at through clustering techniques as is done with Polis [16], or extracted from the type of continuous representations used in Community Notes / Birdwatch [19]. Overall, we view the choice of appropriate representation metric — both for identifying views to inform policies, and evaluating the representativeness of policies — as an open question whose answer may vary depending on the situation.

## A.6 Sample skew

While we employed demographic balancing as part of our sampling procedure, our sample was notably skewed from baselines in the following ways:

- **Ethnicity**—More White and less Hispanic.
- **Education**—More high school only education and less of college or middle school only.
- **Religion**—More non-religious and less Protestant.
- **AI opinion**—More excited about future of AI and less concerned.

<i>Age</i>	US census (18+)	<i>Ours</i>			<i>delta</i>		
		Med	Conflict	Vax	Med	Conflict	Vax
18-24	15	11	11	11	-4	-4	-4
25-34	17	19	20	20	2	3	3
35-44	16	18	18	18	2	2	2
45-54	15	17	18	17	2	3	2
55+	37	35	33	34	-2	-4	-3

<i>Ethnicity</i>	US census	<i>Ours</i>			<i>delta</i>		
		Med	Conflict	Vax	Med	Conflict	Vax
Asian	6	6	6	6	0	0	0
Black	12	15	15	15	3	3	3
Hispanic (Latin)	19	7	7	7	-12	-12	-12
White	58	66	67	67	8	9	9
Mixed	5	5	4	4	0	-1	-1
Other	1	1	1	1	0	0	0

<i>Gender</i>	US census	<i>Ours</i>			<i>delta</i>		
		Med	Conflict	Vax	Med	Conflict	Vax
Male	51	48	48	47	-3	-3	-4
Female	49	50	50	51	1	1	2
Other	0	2	2	2	2	2	2
Prefer not to say	0	0	0	0	0	0	0

<i>Political Party</i>	Gallup 2020	<i>Ours</i>			<i>delta</i>		
		Med	Conflict	Vax	Med	Conflict	Vax
Democrat	31	35	35	35	4	4	4
Republican	25	26	27	28	1	2	3
Independent	41	38	36	36	-3	-5	-5
Other	3	1	2	1	-2	-1	-2

Figure 17: Comparison of demographics in our samples verses baselines (1/2).

<i>Education</i>	US census	<i>Ours</i>			<i>delta</i>		
		Med	Conflict	Vax	Med	Conflict	Vax
Middle school or less	10	1	1	1	-9	-9	-9
High school or GED	28	52	49	50	24	21	22
College/Bachelors degree	45	33	36	36	-12	-9	-9
Masters/PhD or equivalent	13	14	14	13	1	1	0

<i>Religion</i>	US census	<i>Ours</i>			<i>delta</i>		
		Med	Conflict	Vax	Med	Conflict	Vax
Protestant	46	32	31	32	-14	-15	-14
Catholic	21	16	16	16	-5	-5	-5
Mormon	2	1	1	1	-1	-1	-1
Jewish	2	1	2	1	-1	0	-1
Muslim	1	1	1	1	0	0	0
Hindu	1	1	1	1	0	0	0
Other	2	7	8	9	5	6	7
None	24	41	40	39	17	16	15

<i>AI opinion</i>	Pew 2023	<i>Ours</i>			<i>delta</i>		
		Med	Conflict	Vax	Med	Conflict	Vax
More excited than concerned	10	31	30	31	21	20	21
Equally excited and concerned	36	43	44	44	7	8	8
More concerned than excited	52	26	26	24	-26	-26	-28

Figure 18: Comparison of demographics in our samples versus baselines (2/2).

### A.7 Potential overestimation of support

Our participants tended to skew more optimistic toward AI than was observed in a comparable Pew benchmark [20]. At the same time, we found a strong relationship between AI optimism and policy support (figure 19). This means our measurements of policy support may overestimate reality — under certain assumptions,<sup>16</sup> in the most skewed case<sup>17</sup>, the true support could be as much as 10% lower than what we observe.

<i>AI opinion</i>	Pew 2023	Our Sample	<i>Support policy?</i>		
			Yes	Unsure	No
More excited than concerned	10	31	87	8	5
Equally excited and concerned	36	43	78	13	9
More concerned than excited	52	26	54	19	27

Figure 19: Pew data on the percent of Americans with excitement and concern towards AI versus the percent measured in our sample for collective dialogue 3 on *medical advice*. Percent of participants who support the medical advice policy, broken down by their excitement and concern towards AI.

<sup>16</sup>Assuming a) Pew data is perfectly reflective of reality, b) perceptions have not changed since the Pew study, and c) assuming excitement towards AI is the only factor that should be re-weighted for.

<sup>17</sup>CD3 during the *medical advice* policy process.



## A.8 Measuring participant perception

🗨️ Speak

Thank you all so much for your participation and responses.

🗨️ Speak

Before we conclude, we'd like to get your input on how you think this session went.

🗨️ Speak

For context, the purpose of this session was to get public feedback on an initial policy related to AI assistants and providing information and advice related to wars and conflicts. This is just one part of a larger process to create AI policy that reflects the informed consensus of the public. Prior to this session, we ran a session to learn the public's views related to AI providing information and advice related to wars and conflicts, then we identified points of consensus and created an initial policy (the one you saw) based on those. Then we had experts refine that initial policy for things like clarity, consistency, and interpretability. After this session is complete we will be refining the policy we showed you based on your feedback with the goal of making it more representative.

🗨️ Speak

With this in mind, we have a few quick poll questions for you. Each will be a statement, and you will be asked to choose the degree to which you agree with it.

☰ Single-select

"This experience was enjoyable or meaningful."

▼ SHOW OPTIONS

☰ Single-select

"I would trust this process to create a policy which reflects informed public consensus on AI assistants and conflict information."

▼ SHOW OPTIONS

☰ Single-select

"I believe my contributions will find their way to the appropriate place in the final output, and be used appropriately to create a policy which reflects informed public consensus on AI assistants and conflict information."

▼ SHOW OPTIONS

Figure 20: Remesh screenshot showing the context given to participants before asking for their perceptions and the specific statements we asked them to evaluate. For the other policy issues, the string "conflict information" and "wars and conflicts" in the text was replaced with either "medical advice" or "vaccine information."