

# Big Tech-driven deliberative projects

Global Citizens' Assembly Network  
Technical Paper No. 5/2024

**Canning Malkin**

Independent Researcher (United States)

**Nardine Alnemr**

Murdoch University (Australia)



# Preface

Global deliberative democracy was first born as a theoretical experiment. In little over a decade since early publications discussing the norms, conditions, and prospects for such experiment, prototypes for transnational and global deliberation have sprouted in multiple spaces and with varying scale. Most frequently, these processes are initiated by civil society organisations and practitioners who aim to demonstrate that ‘everyday people’ can, and should, have a say in global governance ‘from the bottom-up’. In a few rare cases, supranational organisations have commissioned deliberative processes as well.

What happens when the theoretical and political project of deliberative democracy is instead promoted by private corporations? We posed this question to Nardine Alnemr and Canning Malking, who examine the commissioning context and format for deliberation, purpose and remit, and prospective outcomes of four deliberative processes initiated by tech companies. They deliver much needed documentation of the conditions under which these companies have initiated ‘top-down’ deliberative processes – which differ in several aspects from processes led by civil society and institutions. Unlike governments and civil society, the private sector is not bound to publicising or justifying their decision-making processes to political constituencies, but must instead answer to their investors, boards, and the users of their platforms and products. What will this mean for deliberative democracy?

This technical paper provides valuable insight into the many factors and public arguments surrounding tech-led deliberative processes. As these become more frequent, this insight will become key to understanding to what extent the principles of deliberative democracy, such as equal representation, inclusion of diversity, and public accountability, may (or may not) effectively apply to them.

Melisa Ross

Co-Lead

Global Citizens’ Assembly Network

# Key Findings

This technical paper analyses deliberative projects commissioned by technology companies, such as Meta and OpenAI. We review four tech-commissioned deliberative projects' commissioning contexts, purpose and remit, and outcome:

1. The Royal Society of Arts (RSA) Citizens' Jury on Automated Decision-making Systems (2018), commissioned by DeepMind and implemented by RSA.
2. The Metaverse Community Forum on Bullying and Harassment (2022), commissioned by Meta and implemented by the Stanford Deliberative Democracy Lab in collaboration with the Behavioural Insights Team and Meta.
3. The Democratic Inputs to AI grant program (2023), commissioned by OpenAI and implemented by ten grant-awarded teams.
4. Collective Constitutional AI (2023), commissioned by Anthropic and implemented by the Anthropic and the Collective Intelligence Project (CIP).

The use of deliberative methods to democratise decision-making by technology companies, or to develop prototypes for global deliberation in the tech space, are nascent and experimental. We found that:

1. **These projects cite deliberative democracy as their inspiration.** All four projects cite the literature on deliberative democracy or cases of deliberation in practice as motivation for initiating their own deliberative projects. Deliberation is considered a desirable approach to address the challenge of understanding what the public values in creating rules to govern emerging technologies such as virtual reality spaces or artificial intelligence (AI) chatbots.
2. **Tech companies design deliberative processes with 'global scalability' in mind.** The commissioning companies approached the question of scale either to identify the most feasible deliberation design that would include

and represent participants from around the world or to test AI technology that can facilitate future global citizens' deliberation.

3. **These projects were commissioned after tech-related harms gained significant public attention.** Issues raised in the public sphere include concerns about the use case of AI in healthcare decisions or in content moderation. However, none of the projects' descriptions, justifications, or remit are framed as a direct response to said public concerns. Yet, the commissioning context of each project suggests a link between the motivation to commission and these concerns.
4. **There is a commitment to continue exploring the potential of deliberative projects.** Outcomes of big tech-led deliberative projects did not translate to immediate or concrete changes to tackle the problems submitted to the participants for deliberation. However, all four projects were framed as "seeds" for further experiments incorporating deliberation on how tech companies approach contentious policy issues.
5. **There is a growing network of expertise around how to use deliberation to consider policy options to regulate digital platforms or AI.** Besides organisations involved in the design and implementation of these projects, other groups value citizen deliberation, as well as other participatory approaches, in tech regulation. This opens possibilities for connecting citizen deliberation to tech governance.

For comments and inquiries, please contact Canning Malkin at [canning.malkin@gmail.com](mailto:canning.malkin@gmail.com) and Nardine Alnemr at [nardine.alnemr@murdoch.edu.au](mailto:nardine.alnemr@murdoch.edu.au).

# Background

By the end of 2023, global tech companies such as Meta, OpenAI, and Anthropic commissioned projects inspired by deliberative democracy (henceforth, "deliberative projects"). These projects were commissioned for different purposes, including moderating speech in virtual reality spaces and establishing rules for an AI chatbot. These companies draw on deliberation as a method to include insights from users or participants from the public to democratise the development and management of some technologies.

Pressure on tech companies to make their decision-making more transparent and create effective accountability compatible with fast-paced technological development is not new. There is mounting scrutiny of these companies' role in producing harm, for instance, through violations of workers' rights and the right to privacy,<sup>1</sup> meddling in elections, and the overall decline of the democratic quality of the public sphere.<sup>2</sup> More so, arguments to democratise tech governance call for reigning in these companies' "state-like" powers, e.g., via their influence over policy directions or preferences.<sup>3</sup>

What motivates influential tech companies like Google, Meta and OpenAI to commission projects based on deliberative democracy? What was the purpose of each project? How was deliberation designed and implemented, and what were the outcomes? In this Technical Paper, we describe the role of tech companies in relation to three aspects: the commissioning context, the purpose and remit, and the outcomes of these deliberative projects. The analysis is followed by two considerations to contextualise their role in deliberative

---

<sup>1</sup>L. Giugni (2023, February 08), How tech companies are failing women and social media users—and what to do about it, *The Conversation*. <https://theconversation.com/how-tech-companies-are-failing-women-workers-and-social-media-users-and-what-to-do-about-it-199324>

<sup>2</sup>J. Habermas (2022), Reflections and Hypotheses on a Further Structural Transformation of the Political Public Sphere. *Theory, Culture & Society*, 39(4), 145-171; J. Habermas, J. (2023). A new structural transformation of the public sphere and deliberative politics (C. Cronin, Trans.). *Polity*.

<sup>3</sup>L. Taylor (2021). Public Actors Without Public Values: Legitimacy, Domination and the Regulation of the Technology Sector. *Philosophy & Technology*, 34(4), 897–922; P. Apostolikas (2019), Silicon States: How tech titans are acquiring state-like powers, *Harvard International Review*, 40(4), 18-21.

democracy more broadly based on these projects. The research design, methods and sources used for analysis are included in Appendix 1 on Method and Sources.

To address the question of how tech companies implement deliberation, we showcase different deliberative projects commissioned by AI and tech companies. We selected the following four cases:

1. RSA Citizens' Jury on Automated Decision-making, commissioned by Google DeepMind (2018)
2. Metaverse Community Forum on Bullying and Harassment, commissioned by Meta (2022)
3. Democratic inputs to AI grant program, commissioned by OpenAI (2023)
4. Collective Constitutional AI, commissioned by Anthropic (2023)

Appendix 2 presents a table summarising findings across the four cases.

# Findings

Hosting deliberative forums with consumers is not as common as hosting deliberative forums with citizens.<sup>4</sup> Deliberative forums with consumers present opportunities for global deliberation when companies with a global consumer base are involved. The cases we discuss here are of high import as the companies hold considerable global influence and can be regarded as norm-setters for future deliberative processes commissioned by private companies, just as much for the AI industry where these cases concentrate.

The projects surveyed here are nascent, and so is the research. Therefore, our objective is twofold: to provide a snapshot of these projects from conception to completion and to contextualise the implications of these projects in deliberative democracy more broadly. Our analysis focuses on three dimensions in tech-commissioned deliberative projects:

1. The commissioning context<sup>5</sup>
2. The purpose and remit
3. The outcomes

The analysis and insights presented below are limited by two caveats. First, these processes are incipient in the industry and worldwide. The following discussion is, therefore, neither extensive nor exhaustive. As these experiments grow in numbers, we focus on documenting salient features of these examples as norm-setters for future tech-commissioned deliberative projects. Second, our analysis does not include an assessment of the deliberative quality of these processes. This is a research agenda in its own right and demands a different approach, research design, and access to data. In this technical paper, we concentrate on the context and design features of these first experiments in tech companies to understand how they adopt the principles and different designs of deliberation.

---

<sup>4</sup> See C. Hendriks (2009), *Deliberative governance in the context of power*, *Policy and Society*, 28(3), 173-184. <https://www.tandfonline.com/doi/pdf/10.1016/j.polsoc.2009.08.004>

<sup>5</sup> Based on C. Malkin et al. (2023). *Agenda-setting in transnational and global citizens' assemblies*. GloCAN Technical Paper No. 2/2023. <https://glocan.org/wp-content/uploads/2024/01/Technical-Paper-2-2023-Malkin-et-al.pdf>

## 1. The commissioning context

Why were each of these projects commissioned? What is the relevant policy concern? Who was involved in the design and implementation? What was the selected deliberation design and why? We analyse these questions for each case, explaining how decisions were made considering three intersecting opportunities. Briefly, these are:

- **Values alignment.** The commissioners and collaborators share an interest in citizen or user input and engagement to align their technology development with public values, e.g. with respect to what rules AI-chatbots should observe in communication or rules to enforce in regulating private virtual spaces.
- **Political opportunity.** The commissioners tend to host these projects in conversation with an ongoing policy or regulatory concern that has gained public attention.
- **Strategic opportunity.** The commissioners and collaborators (including designers and implementers) use these projects to position themselves in the market, and to identify potential collaborations and pathways for public engagement and participation.

We present the commissioning context and these three opportunities for each case in turn.

### 1.1. *RSA Citizens' Jury on Automated Decision-making: Political and values alignment following concerns over AI advancement*

DeepMind commissioned a citizens' jury on automated decision-making following an investigation by the UK's Information Commissioner's Office in July 2017, where the company failed to comply with the Data Protection Act of 1998 in processing the personal health data of 1.6 million patients.<sup>6</sup> The violations sparked calls for stronger accountability measures, specifically to empower citizens to decide how algorithmic and automated systems are used in

---

<sup>6</sup> See Information Commissioner's Office (July 2017), RFA0627721 – provision of patient data to DeepMind. <https://ico.org.uk/media/action-weve-taken/undertakings/2014353/undertaking-cover-letter-revised-04072017-to-first-person.pdf>



decisions that affect them.<sup>7</sup> Therefore, this backdrop brought forward an opportunity for values alignment and informing policy.

The Royal Society for the Arts (RSA) was commissioned to implement the project. The project involved two deliberations: a citizens' jury and an enclave deliberation (i.e., deliberation amongst a homogenous group).<sup>8</sup> The citizens' jury recruited 29 participants selected to "broadly reflect the makeup of the population of England and Wales" who met in-person for four days over five months.<sup>9</sup> The RSA collaborated with a market research company for participant recruitment and an independent organisation, Deliberate Thinking, for jury design and facilitation. The lead facilitator, Diane Beddoes from Deliberate Thinking, also trained support facilitators from the RSA team. The RSA hosted the deliberation enclave for individuals from communities disproportionately affected by automated decision-making systems: young men from Black, Asian and Minority Ethnic (BAME) background.<sup>10</sup> The mixed format of citizens' jury and deliberation enclave tap into the opportunities for (1) aligning with public values and those of disproportionately affected communities and (2) platforming these voices in the ongoing policy debate about remedies for algorithmic and AI harms to society.

### 1.2. *Metaverse Community Forum on Bullying and Harassment: Strategic, political and values alignment opportunities for Meta*

The Metaverse Community Forum on Bullying and Harassment was an online deliberative event in which 6,488 participants from 32 countries, grouped in nine geographical regions across 19 language clusters. They were tasked with discerning adequate rules to limit bullying and harassment in private spaces in

---

<sup>7</sup> See, for example, H. Shah (2017), The DeepMind debacle demands dialogue on data, Nature.

<https://www.nature.com/articles/547259a>

<sup>8</sup> The RSA (2019), Democratising decisions about technology: A toolkit (Report).

<https://www.thersa.org/globalassets/reports/2019/democratising-decisions-tech-report.pdf>

<sup>9</sup> Deliberate Thinking website <https://deliberatethinking.co.uk/about/>

<sup>10</sup> The RSA report does not provide details on which communities the participants are from.

BAME was the preferred term by the government at the time. The UK government advised it is no longer using BAME or BME in April 2022. <https://equalities.blog.gov.uk/2022/04/07/why-we-no-longer-use-the-term-bame-in-government>

the Metaverse (Meta’s virtual reality social media).<sup>11</sup> This Community Forum was a collaboration between Meta, the Stanford Deliberative Democracy Lab, and the Behavioural Insights Team. It is the first of a series of forums that seek to engage social media users. Meta commissioned the Stanford Deliberative Democracy Lab to conduct a Deliberative Poll after hosting a transnational digital citizens’ assembly on climate misinformation.<sup>12</sup> The Lab celebrates this experiment as the first global Deliberative Poll, successfully recruiting more participants than the expected 5,900.<sup>13</sup>

According to our interviewee, the strategic opportunity for Meta was to identify an effective method to engage with their global user base in a process that would produce data on specific issues. The primary decision-makers in the design and implementation of the process were representatives from Meta, the Deliberative Democracy Lab, the Behavioural Insights Team,<sup>14</sup> and an Advisory Committee, selected by Meta. Included in the Advisory Committee’s responsibilities were “vetting [sic] the briefing materials for the deliberations and providing [sic] many of the experts for the plenary session.”<sup>15</sup> The Lab also partnered with multiple market research companies for recruitment and used a suite of AI technologies, including an AI-moderator.

Content moderation has long been a contentious question for Meta. Setting “bullying and harassment” as the topic for deliberation could be seen as a political opportunity for the company to address wide-ranging concerns about online safety on its platforms, including increasing concern over sexual assault

---

<sup>11</sup> J. Fishkin et al. (June 2023), Meta Community Forum: Results Analysis, Stanford Deliberative Democracy Lab. <https://cddrl.fsi.stanford.edu/publication/metaverse-community-forum-results-analysis>

<sup>12</sup> E. Costa (November 2022), Deliberative Democracy in action: A closer look at our recent pilot with Meta. <https://www.bi.team/blogs/deliberative-democracy-in-action/>.

<sup>13</sup> Deliberative Democracy Lab (June 2023), Results of First Global Deliberative Poll<sup>®</sup> Announced by Stanford’s Deliberative Democracy Lab, Stanford University. <https://cddrl.fsi.stanford.edu/news/results-first-global-deliberative-pollr-announced-stanfords-deliberative-democracy-lab>

<sup>14</sup> A company that specializes in applying behavioural insights to systems and policies.

<sup>15</sup> J. Fishkin et al. (June 2023), Meta Community Forum: Results Analysis (see page 4). [https://fsi9-prod.s3.us-west-1.amazonaws.com/s3fs-public/2023-06/public\\_meta\\_community\\_forum\\_final\\_report\\_-\\_stanford\\_ddl\\_1.pdf](https://fsi9-prod.s3.us-west-1.amazonaws.com/s3fs-public/2023-06/public_meta_community_forum_final_report_-_stanford_ddl_1.pdf).

and harassment experienced in the virtual Metaverse, especially against children.<sup>16</sup>

As for values alignment, according to Meta’s Transparency website, the aim of the Community Fora is to “give people a greater voice in the governance of our platforms and development of the technologies that affect them.”<sup>17</sup> Public input and deliberation complements Meta’s existing governance structure, which is the Oversight Board.

### 1.3. *Democratic Inputs to AI: Values alignment and strategic opportunities to develop technology for global deliberation*

A few days after the release of OpenAI’s fourth version of its AI chatbot (ChatGPT-4), the company’s CEO Sam Altman was asked in an interview about guardrails for AI safety against discriminatory and hateful content. In response, Altman argued that the best approach to align the values governing AI models while also balancing competing values would be a global deliberative process:

My dream scenario, and I don’t think we can quite get here, but like, let’s say this is the platonic ideal, and we can see how close we get, is that *every person on earth* would come together, have a really thoughtful, *deliberative conversation* about where *we want to draw the boundary on this system*. And we would have something like the *US Constitutional Convention*, where we debate the issues and we look at things from different perspectives and say, well, this will be good in a vacuum, but it needs a check here.<sup>18</sup> (emphasis added)

In reaction to this statement, a few followers of both global deliberation and tech regulation have noted that a proof-of-concept exists: The Global Assembly

---

<sup>16</sup> See, A. Phippen (March 2022), Protecting children in the metaverse: it’s easy to blame big tech, but we all have a role to play, LSE Blog, <https://blogs.lse.ac.uk/parenting4digitalfuture/2022/03/23/metaverse/>; Z. Smith (February 2024), It Takes a Village to Protect Children in the Metaverse: U.S. Global Engagement Initiative, Carnegie Council for Ethics in International Affairs.

<https://www.carnegiecouncil.org/media/article/protect-children-metaverse>

<sup>17</sup> Meta (April 2024), Governance Innovation.

<https://transparency.meta.com/governance/innovation/>.

<sup>18</sup> Lex Fridman (26 March 2023), Sam Altman: OpenAI CEO on GPT-4, ChatGPT, and the Future of AI | Lex Fridman Podcast #367, YouTube. [https://www.youtube.com/watch?v=L\\_Guz73e6fw](https://www.youtube.com/watch?v=L_Guz73e6fw)

on the Climate and Ecological Crisis (a global deliberative process), and the Alignment Assemblies<sup>19</sup> hosted by the Collective Intelligence Project (CIP) which provide examples of what Altman proposes: a deliberative process about “where we want to draw the boundary on this system”.<sup>20</sup> Details about developments in this conversation are not public, but two months later, OpenAI’s non-profit launched its Democratic Inputs to AI program in May 2023. The project would give ten successful applicants a grant of \$100,000 each to design a prototype of a “democratic process”, which would test models of governance for artificial general intelligence (AGI).

The opportunity for values alignment also drove OpenAI’s agenda-setting for this project: creating prototypes for a deliberative process defined as an event in which “a broadly representative group of people exchange opinions, engage in deliberative discussions, and ultimately decide on an outcome via a transparent decision-making process.” In announcing this project, OpenAI made a statement that with the wide-ranging effects of AI technology on society, “AGI should benefit all of humanity and be shaped to be as inclusive as possible.” Nonetheless, this announcement took place at the same time as some OpenAI employees left the company, citing concerns for how OpenAI is approaching AI alignment.

Altman’s initial thoughts on a global deliberative governance structure on AI suggest that, to a certain degree, the company has an alignment of values. While this does not equate to company policy, in its call for applications, OpenAI said, “Technology shapes the lives of individuals, how we interact with one another, and how society as a whole evolves. We believe that decisions about AI should be shaped by diverse perspectives reflecting public interest.”<sup>21</sup>

As for the strategic opportunity, OpenAI’s grant program created connections to potential collaborators. Ten projects from twelve countries were selected from a pool of nearly 1,000 projects. The selection of the grant winners has the

---

<sup>19</sup> The Alignment Assemblies aim to generate public insights on how AI development can support “the public good”. For more information, see The Collective Intelligence Project (November 2023), Alignment Assemblies: Nine Months In. <https://cip.org/blog/alignment-assemblies-nine-months-in>

<sup>20</sup> Ibid; also see for instance <https://twitter.com/RichDecibels/status/1641118020918558725>

<sup>21</sup> Direct quotes in this segment from W. Zaremba et al. (May 2023), Democratic inputs to AI. <https://openai.com/blog/democratic-inputs-to-ai>.

potential to influence the uptake of these prototypes in transnational or global deliberation in the future.

#### *1.4. Collective Constitutional AI: Opportunity for values alignment with public preferences and strategic opportunity in the race to lead*

As a result of a division amongst OpenAI employees over a few issues, one of which concerns future directions to govern their AI models, 11 employees left the company and founded Anthropic to actualise their vision for “harmless and helpful AI”. They call their approach Constitutional AI, to train a public interest-aligned AI. In December 2022, the company argued that this approach can be democratised to include input from “diverse stakeholders”.<sup>22</sup>

In September 2023, Anthropic’s CEO Dario Amodei attributed this concern over AI assistants aligned with values held by the public as the key driver for leaving OpenAI.<sup>23</sup> Considering the timing of this public statement overlapped with OpenAI’s grant, Amodei may have also attempted to race OpenAI to claim the space of “alignment” to be associated with their company, not their rival. Moreover, the language used is similar to what Altman said in his interview, where he also expressed interest in a global deliberative process that borrows on the U.S. “constitutional review” model, implying that the AI chatbot’s actions can be reviewed against its constitutional rules. Our interpretation of the timing and the similarity in framing is that Anthropic’s motivations may be explained as part of a competition dynamic to establish leadership in using democratic processes to govern AI development.

To further demonstrate the strategic opportunity to demonstrate their leadership, Anthropic organised an online deliberative public-input experiment in collaboration with the Collective Intelligence Project (CIP). In this experiment, 1000 participants were selected as a representative sample of the U.S. were asked to vote on 40 statements or more and they also had the option to create statements of their own. Statements that received consensus

---

<sup>22</sup> Anthropic (December 2022), Constitutional AI: Harmlessness from AI Feedback, Policy Memo.

<https://cdn.sanity.io/files/4zrzovbb/website/7512771452629584566b6303311496c262da1006.pdf>

<sup>23</sup> Fortune Magazine (September 2023), Why Anthropic's Founder Left Sam Altman’s OpenAI, YouTube. <https://www.youtube.com/watch?v=5GtVrk00eck>

were then taken by Anthropic as “principles” which they curated into a “public constitution”.<sup>24</sup> Anthropic researchers then sought to assess the extent to which their in-house constitution matches or differs from public values – they found a 50% overlap. Discussions were moderated by the CIP team. While collaborations with CIP presented a strategic opportunity to both OpenAI and Anthropic, it is hard to discern if the interest in deliberative democracy has motives beyond competing with OpenAI.

## **2. The purpose and remit**

In this section, we focus on the announced purpose and remit of each deliberative project: what did they seek to achieve by adopting a deliberative project? The declared purpose directly shaped the scope of each project: the subject of discussion for participants and users (DeepMind-RSA, Meta-Stanford Deliberative Democracy Lab, and Anthropic-CIP) or the challenge for innovations to create prototypes for global deliberation (OpenAI vis-à-vis the 10 grant recipients).

### *2.1. RSA Citizens’ Jury: A toolkit for participatory AI governance*

The purpose of this project was to generate citizens’ perspectives on AI regulation to complement views in policymaking spaces and in conversations with the industry. Hence, jurors deliberated on the remit: “Under what conditions, if any, is it appropriate to use an automated decision system?”. Three use cases were proposed for consideration: job recruitment, healthcare, and criminal justice. The topics were inspired by policy concerns over using automated decision-making in 2018. For example, software like COMPAS, used to aid in recidivism judgment in the U.S. criminal justice system, was proven to be biased against Black defendants.<sup>25</sup>

---

<sup>24</sup> Direct quotes in this segment from Anthropic (October 2023), Collective Constitutional AI: Aligning a Language Model with Public Input. <https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input>

<sup>25</sup> Direct quotes in this segment from J. Angwin et al. & ProPublica. (May 2016). Machine Bias: There’s software used across the country to predict future criminals. And it’s biased against blacks, ProPublica. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

DeepMind was not involved in setting the purpose or the remit. The RSA research team, in consultation with an Advisory Board and with input from 24 expert witnesses, decided on it. The RSA also selected the Board, which included experts in AI/automated systems (in business, government, academia, and civil society) as well as practitioners in deliberative engagement. Output of the research (the jury, enclave and YouGov surveys of non-participants) were used to develop a toolkit for “organisations seeking to deploy their ethical processes around the proliferation of AI.”<sup>26</sup>

## 2.2. *Metaverse Community Forum: Moderation rules for bullying and harassment*

The Community Forum was initiated to identify user preferences in response to the problem of bullying and harassment in “private virtual worlds,” which are spaces in the Metaverse that are created and moderated by individual users rather than Meta. Representatives from the Deliberative Democracy Lab, Meta, the Behavioural Insights Team, and the Advisory Committee, collectively decided on the question posed to participants in the project according to an interview partner.

Participants in this process were presented with a specific remit: “How to regulate bullying and harassment in virtual reality, particularly in the new private or ‘members-only’ social virtual reality spaces that are being created in the Metaverse?.” The question was limited to “bullying and harassment” and excluded “other abusive or illegal online behavio[u]rs”. The Lab’s team developed different iterations of the remit to clarify to the participants the narrowness of this scope (e.g., “to what extent should the platform owners stay out, since these ‘members-only’ spaces are not public [...] or to what extent do the platform owners, such as Meta, have a responsibility to act to protect against bullying and harassment?”).<sup>27</sup>

## 2.3. *Democratic input to AI: Prototypes for deliberative AI governance*

---

<sup>26</sup> See The RSA, Democratising decisions about technology. Information about the research team is on page 3 and members of the advisory board on page 52.

<sup>27</sup> Direct quotes in this segment from J. Fishkin et al. (June 2023), Meta Community Forum, page 4.

OpenAI sought “innovative prototypes” of deliberation from across the world in their grants program. The scope was broadly around designing and developing prototypes for global deliberation to be used in AI governance. While commissioners typically can define the subject of deliberation, OpenAI in this project set the parameters for design. To guide the applicants further, OpenAI presented them with a list of relevant policy questions to consider. However, key issues to AI governance, such as “transparency” and “training data copyrights”, were not included in the list.<sup>28</sup>

#### *2.4. Collective Constitutional AI: Comparative AI constitution-making*

Anthropic’s product is Claude, a large language model (LLM) AI chatbot similar to OpenAI’s ChatGPT. Claude’s responses to queries from users are restrained by a rules-based system which Anthropic refers to as “Constitutional AI”, inspired by how constitutions govern relations in a state.

Collaborating with the Collective Intelligence Project, the purpose of Anthropic’s deliberative project was to compare two versions of Constitutional AI: one which is created by Anthropic’s team and the other, Public Constitution for AI, curated by a sample of the U.S. through a deliberative process. The aim of the deliberation was to capture a range of public values and consensus on these, in order to curate "principles" that would be incorporated into the Public Constitution for AI. According to Anthropic, the deliberation involved nearly 1,000 participants who voted on their preference for rules in Anthropic’s AI Constitution and made 1,127 more statements to be considered by other deliberators for inclusion in the Public Constitution.<sup>29</sup>

### **3. The outcomes**

These projects are in early stages so it remains unclear how the outputs of each deliberative project will be used by the commissioning companies. One consistent outcome across these projects is the companies’ interest in future iterations. These take the form of either inspiring further efforts to link the

---

<sup>28</sup> L. Gissen (August 2023), Don’t Use Deliberative Democracy to Distract from Regulation. <https://democracy-technologies.org/opinion/dont-use-deliberative-democracy-to-distract-from-regulation/>

<sup>29</sup> Anthropic (October 2023), Collective Constitutional AI.



outputs of deliberation to policymaking or for the companies themselves to continue exploring how deliberation might be integrated into their own policymaking. For instance, the RSA anchored its project with DeepMind in its advocacy for citizen engagement in ethical AI governance. The companies involved in the other projects might not be committed to such advocacy to link with national or international policymaking. Their deliberative projects instead would be instrumental to exploring moderation policy options (Meta and Anthropic) or prototyping that advance their own technologies (OpenAI). In this section, we review the outcomes in each case.

### *3.1. RSA Citizens' Jury: Aspiration for deliberation-centric engagement with business, government, and civil society*

In their report, the RSA aimed to host a public event with businesses, government, and civil society to respond to citizens' insights.<sup>30</sup> DeepMind limited its role to commissioning and funding, unlike the companies commissioning the other projects discussed here. The RSA complemented its work on this project with broader advocacy for the engagement of citizens and experts in conversations about ethical roles for AI in society,<sup>31</sup> which appear to have resonated with government institutions. A year after the project in 2019, two more citizens' juries on AI governance took place in the UK, commissioned by National Institute for Health Research (NIHR) Greater Manchester Patient Safety Translational Research Centre (PSTRC) and the Information Commissioner's Office and implemented by the Jefferson Center.<sup>32</sup> Moreover, the 2018 RSA jury was cited as a positive example of public engagement in response to racialised harm evidenced in AI recruitment systems.<sup>33</sup>

---

<sup>30</sup> No information was found about this event specifically. However, the RSA did have a series supported by DeepMind: You and AI. <https://royalsociety.org/-/media/policy/projects/machine-learning/you-and-AI-summary.pdf?la=en-GB&hash=98E2FF41B07E0CD1E6F3F2684C27C7A5>

<sup>31</sup> See B. Balaram et al. (May 2018), Artificial Intelligence: real public engagement, RSA Reports, Medium. <https://medium.com/rsa-reports/artificial-intelligence-real-public-engagement-6b0fd073e2c2>; Singh (October 2019), We need to talk about artificial intelligence, The RSA website. <https://www.thersa.org/blog/2019/10/talk-about-artificial-intelligence>.

<sup>32</sup> Centre for New Democratic Processes (2019), Citizens Juries on Artificial Intelligence. <https://jefferson-center.org/citizens-juries-artificial-intelligence/>

<sup>33</sup> R. Booth (October 2019), Unilever saves on recruiters by using AI to assess job interviews, The Guardian. <https://www.theguardian.com/technology/2019/oct/25/unilever-saves-on-recruiters-by-using-ai-to-assess-job-interviews>

### 3.2. *Metaverse Community Forum: Further experiments with Deliberative Polls*

The Stanford Deliberative Democracy Lab designed the Poll and their research to make participant insights actionable for Meta or “other platform owners”.<sup>34</sup> Since the completion of the Bullying and Harassment in the Metaverse Community Forum, Meta has launched a website<sup>35</sup> which will describe their Community Fora and serve as the home of publications from the company about its deliberative events.

Like other cases discussed here, Meta made no direct commitment to uptake recommendations from the deliberations. Instead, the company affirmed its commitment to continue experimenting to identify the most feasible democratic process for including users in governance. The company’s recent Deliberative Poll on Generative AI also demonstrates this interest in testing the method in Meta’s engagement with its global user base.

### 3.3. *Democratic Inputs to AI: Further experiments with deliberation for values alignment*

OpenAI made clear from the beginning of their Democratic Inputs project the outcomes would not have an immediate impact but would feed into further experiments with developing technology to facilitate global deliberation. OpenAI’s announcement of the project explains:

[w]hile these initial experiments are not (at least for now) intended to be binding for decisions, we hope that they will explore decision-relevant questions and build novel democratic tools that can more directly inform decisions in the future.<sup>36</sup>

This opens a question about the possibility of linking tech-commissioned deliberative processes with participants from the public to policymaking spaces. We expand on this point below, in the discussion on contextualising tech-commissioned projects in deliberative democracy.

---

<sup>34</sup> J. Fishkin et al., Meta Community Forum.

<sup>35</sup> Meta Transparency Center, Governance Innovation (Updated April 2024), <https://transparency.meta.com/governance/innovation/>.

<sup>36</sup> W. Zaremba et al., Democratic inputs to AI.

At the end of the grant round, OpenAI announced the creation of a Collective Alignment Team “consisting of researchers and engineers,”<sup>37</sup> tasked with continuing the effort to design systems that will incorporate public input into AI governance. The Team would also be responsible for engagement with external organisations and grantees to incorporate their models into the company’s “steering.”

### 3.4. Collective Constitutional AI: Aspiration to motivate other AI teams

In terms of outcomes, Anthropic frames their deliberative project with CIP as a research endeavour but suggest that other teams may find deliberative democracy relevant in AI development. Therefore, the outcomes are limited to signalling leadership in deliberative engagement against their competitors and making statements about inspiring interest in these experiments. For example, Anthropic’s blog reiterates the leadership claim:

We believe that our work *may be one of the first instances* in which members of the public have *collectively directed* the behavior of a large language model through written specifications *via an online deliberation process*. We hope that sharing our very preliminary and imperfect findings sooner rather than later will *help others interested in democratic inputs to AI* to learn from our successes and failures. (emphasis added)<sup>38</sup>

As for experimentation, Anthropic’s policy memo reads:

This effort was an experiment in developing a more democratic process and methodology for training AI models. *It leaves room for future iteration in several areas:* the individuals and communities selected to participate in public input processes, the curation and composition of resulting constitutions, and the evaluations used to assess downstream model behavior, among others. (emphasis added)<sup>39</sup>

---

<sup>37</sup> OpenAI, Democratic inputs to AI grant program: lessons learned and implementation plans.

<sup>38</sup> Anthropic (October 2023), Collective Constitutional AI.

<sup>39</sup> Anthropic (2023), Collective Constitutional AI: Aligning a Language Model with Public Input, Policy Memo. [https://www-cdn.anthropic.com/b43359be43cabdbe3a8ffd60ea8a68acf25cb22e/Anthropic\\_CollectiveConstitutionalAI.pdf](https://www-cdn.anthropic.com/b43359be43cabdbe3a8ffd60ea8a68acf25cb22e/Anthropic_CollectiveConstitutionalAI.pdf)

## 4. Conclusion

Tech companies' use of deliberative methods is in the early and experimental stages. Thus far, outcomes are limited to inspiring further experiments and rely on whether any of the companies will set an example by committing to be accountable to recommendations from these deliberations. Based on our observation of the time of commissioning and announcing the deliberative projects in 2023, we think considerations for higher commitment are likely constrained by weighing the impact it may have on competition and policy options.

Finally, researchers, practitioners, policymakers, and companies should carefully evaluate the purpose and outcomes of these commissioned deliberative projects to avoid the risk of "participation washing".<sup>40</sup> Followers of developments in digital democracy, like Laura Gissen, warned about the conflict of interest at the heart of these processes where deliberation can be co-opted in lobbying strategies.<sup>41</sup> In the discussion below, we emphasise that while there is room for experimentation, multiple stakeholders should also consider the implications of these deliberative experiments for the public accountability of tech companies.

---

<sup>40</sup> M. Sloane et al. (2020). Participation is not a design fix for machine learning. <https://arxiv.org/ftp/arxiv/papers/2007/2007.02423.pdf>

<sup>41</sup> L. Gissen, Don't Use Deliberative Democracy to Distract from Regulation.

# Recommendations

In the previous section, we surveyed the commissioning context, purpose and remit, and outcomes of tech-commissioned deliberative projects to explain the uptake witnessed in 2023. We anticipate that these experiments will continue to grow, and new deliberative projects will be implemented by these and other tech companies.

In consequence, we highlight three considerations for future iterations and research:

1. **For commissioning companies and implementers: The purpose of deliberative projects should disentangle democratizing tech development from developing technologies that can aid global deliberation.**

We particularly see this in OpenAI's Democratic Inputs to AI project where the developing a technology that would enable global deliberation about AI governance is seen as a precursor. Using deliberation as a method for user input to decision-making or input to technology development is valid but should be seen as separate pursuits, and they present different ramifications for deliberative democracy. For the former, this would require companies to acknowledge the possibilities and limits of incorporating user input. As for the latter, it requires engagement with other actors aspiring to create citizens' assemblies at the global level.

2. **For commissioners and implementers: Transparency about the outcome and potential impact of the deliberative projects is key.**

Commissioners and implementers ought to clearly communicate to participants the extent to which their input and recommendations will have an influence on a company's decision regarding the remit and issue submitted to deliberation. This finding echoes points raised in GloCAN

Technical Paper 1/2024 on the backstage governance of citizens' assemblies.<sup>42</sup>

**3. For policymakers and tech companies: Build on recommendations from deliberative engagement on AI governance.**

As highlighted in the outcomes of the DeepMind commissioned RSA citizens' jury which was followed by two citizen deliberation processes in the UK on AI, there is potential to continue building on the principles, preferences, and arguments identified in deliberative processes so far, rather than starting anew with every new iteration or participatory experiment.

---

<sup>42</sup> R. Dean et al. (2024), Spotlighting the backstage governance of citizens' assemblies, GloCAN Technical Paper No. 1/2024. <https://glocan.org/wp-content/uploads/2024/01/Technical-Paper-1-2024-Dean-et-al.pdf>

# Outlook

## **Contextualising tech-commissioned deliberative projects in deliberative democracy.**

In this paper, we presented factors that likely shaped tech companies' interest in deliberative projects. The significance of this work is to document these projects soon after their conclusion and to present a power-sensitive account considering the global influence of the commissioning companies on communication, deliberation, and democracy. Similar to the authors' findings in GloCAN's Technical Paper No. 2/2023 on agenda-setting,<sup>43</sup> identifying the context, deliberation design, purpose and remit for each project manifests the power of commissioners and implementers to shape the boundaries for discussions and outcomes.

In this section, we briefly discuss two questions about the future development of these projects and how to contextualise the role of tech companies in deliberative democracy. Finally, we present two suggestions for future research based on the four cases analysed above.

### **I. Will deliberative democracy shape tech companies, or will tech companies shape deliberation?**

As explained earlier for each case study, deliberation was not implemented in the same way across deliberative projects. Projects commissioned by Google DeepMind, Meta, and Anthropic used deliberation to get insights from participants on a specific policy issue or concern. This is aligned with how deliberative processes, with their varied formats, are commissioned by other actors. OpenAI, in contrast, used the Democratic Inputs to AI grant program to develop prototypes that would enable global deliberation, i.e., technology that would embody inclusion and sortition.

---

<sup>43</sup> C. Malkin et al., Agenda-setting in transnational and global citizens' assemblies.

This creates an opportunity to critically reflect on the emerging role of tech companies in deliberative democracy. On the one hand, this can involve interrogating how technology may aid in global citizen deliberation. For example, if Meta continues to experiment with Deliberative Polls, commissioners, designers, implementers, and observers can then discern the costs, logistics, resources, and technology needed for global deliberative processes. On the other hand, if the development of technology that aid deliberation is done without active engagement with deliberation designers, implementers, or facilitators can reflect values or priorities that are based on technological logics rather than experience.

A final consideration is whether, and when, the commissioning companies make more concrete commitments to the outcomes of their deliberative projects. Doing so would be norm-setting considering the global influence these companies have on online communication and deliberation more broadly. While these projects are nascent, it is important not to underestimate the interest of tech companies in deliberative democracy.

## **II. What are the implications of tech-led deliberation for democratic accountability?**

The commissioning contexts for these forums have been, to varying degrees, spurred by scrutiny of the companies' and the industry's policies or products (see for instance sections 1.1. and 1.2. above). Three out of the four projects centred on AI governance, a topic under growing scrutiny, politically and socially.<sup>44</sup> Therefore, analysis of the commissioning context and motivation to commission should document to the relevant political concerns or policy issues.

While government-commissioned deliberative events can respond to salient issues, governmental bodies are responsible for policy decisions in nearly all areas. In contrast, private companies do not

---

<sup>44</sup> S. Mukherjee (May 2023), Top AI CEOs, experts raise “risk of extinction” from AI.  
<https://www.reuters.com/technology/top-ai-ceos-experts-raise-risk-extinction-ai-2023-05-30>



have the same obligation to respond. This has several consequences. First, initiating public deliberation as a reaction to social, political and cultural backlash can impose a ceiling on deliberation before an event has been commissioned. Additionally, because private companies do not have the same duty of transparency to the public as governments do, these deliberative projects risk falling by the wayside as examples of stakeholder engagement. While our case studies generally have some level of outward commitment regarding the products of deliberation, this could happen in the future. We present this as an area for continued research; as the impacts of these deliberative events become clearer, analysis of how wide, deep, and consistent their reach will be increasingly necessary.

In another respect, the levels of a company's involvement in the deliberative project should match accountability for the outcomes. Google DeepMind restrained itself as merely funding and commissioning, giving the RSA autonomy to execute the project. Meta, OpenAI and Anthropic, on the other hand, were directly engaged in different stages of the project. Yet, considering these are experimental projects, expectations of accountability are difficult to identify. Strengthening accountability by linking these deliberative projects to enforceable policymaking is possible.

### **Potential for research and future collaborations.**

From our analysis of the commissioning contexts, purpose, and outcome, we conclude that deliberative projects commissioned by technology companies demonstrate possibilities for collaborations between the companies and groups that design and implement deliberative processes. Through these experiments and collaborations lies the potential to define how deliberation may influence decisions made within these companies to better reflect insights from the public. Despite this, the evidence is limited to make substantive assessments about these projects' potential impact on democratising tech governance or the potential of AI technology to facilitate global deliberation.

In light of the previous questions and the findings from the four case studies, future research agenda could focus on:

1. **Policy debates or concerns in the backdrop for tech-commissioned deliberative projects.** With the companies' power and influence, we also ought to ask: what role can these projects inspired by deliberative democracy empower citizens in tech governance, if at all? With the above considerations in mind, research ought to also document the extent to which tech-commissioned deliberative projects contribute to better global deliberation or relegate it to consultative episodes with users.
2. **Other actors who shape these deliberative projects.** While the focus herein has been on the commissioning tech companies, there are other actors who can substantively shape the direction of these experiments. Analysis in the future could consider comparing efforts led by civil society organisations using deliberation for tech governance and AI oversight. Active organisations in the field include the Ada Lovelace Institute,<sup>45</sup> the Collective Intelligence Project,<sup>46</sup> the RSA,<sup>47</sup> and AlgorithmWatch.<sup>48</sup> More crucially, insights from these cases and beyond present an opportunity to synthesize governance recommendations. This can take the form of systematic analysis of the content of these deliberations about ideals and principles to observe in AI governance.

---

<sup>45</sup> L. Groves (December 2023), Going public: Exploring public participation in commercial AI labs (Report), Ada Lovelace Institute. <https://www.adalovelaceinstitute.org/report/going-public-participation-ai/>

<sup>46</sup> See The Collective Intelligence Project, Alignment Assemblies <https://cip.org/alignmentassemblies>

<sup>47</sup> See The RSA, Tech and Society <https://www.thersa.org/projects/archive/economy/tech-and-society>

<sup>48</sup> See AlgorithmWatch and AI Forensics, ChatGPT and Co: Are AI-driven search engines a threat to democratic elections? <https://algorithmwatch.org/en/bing-chat-election-2023/>

# Appendix

## Appendix 1. Method and Sources

The data collection and analysis for this technical paper started in October 2023 and ended in January 2024. We have excluded from our analysis the Metaverse Community Forum Deliberative Poll on Generative AI (October 2023).<sup>49</sup> The first publication with details about this Poll was released in late January 2024, which coincided with the time we finalized our analysis of the four cases.<sup>50</sup>

Between November-December 2023, we sent interview requests to relevant stakeholders who declined to be interviewed, except for one. Insights from this single interview are used for contextual clarification. In re-thinking the structure for this technical paper, we have added the RSA project to the case studies and analysed publicly available sources such as the companies' public announcements of the project and outcomes and reports on the project by the implementing partners.

To supplement the analysis, we also reviewed publicly available interviews with tech company CEOs and coverage of the projects by specialised outlets such as Democracy Technologies who observed and reported on tech-commissioned deliberation. During the final revisions, we have contacted the RSA's archive team to clarify details in the jury's documentation. Publicly available documents about the four case studies are catalogued in the following list:

Case study	Title	Document type	Link
	Democratising decisions about technology: A toolkit	Report	<a href="https://www.thersa.org/reports/democrati">https://www.thersa.org/reports/democrati</a>

<sup>49</sup> See N. Clegg (June 2023), Bringing people together to inform decision-making on generative AI. Meta <https://about.fb.com/news/2023/06/generative-ai-community-forum/>; Stanford Deliberative Democracy Lab (October 2023), Meta Community Forum on Generative AI.

<sup>50</sup> G. Wetherall-Grujić (January 2024), Meta experiments with public consultation on AI, Democracy Technologies: <https://democracy-technologies.org/participation/meta-community-forums-generative-ai/>

RSA Citizens' Jury on Automated Decision-making Systems (2018)			<a href="#">sing-decisions-technology-toolkit</a>
	We need to talk about artificial intelligence	Blog post	<a href="https://www.thersa.org/blog/2019/10/talk-about-artificial-intelligence">https://www.thersa.org/blog/2019/10/talk-about-artificial-intelligence</a>
	Results of First Global Deliberative Poll Announced by Stanford's Deliberative Democracy Lab	News update	<a href="https://cddrl.fsi.stanford.edu/news/results-first-global-deliberative-poll-announced-stanfords-deliberative-democracy-lab">https://cddrl.fsi.stanford.edu/news/results-first-global-deliberative-poll-announced-stanfords-deliberative-democracy-lab</a>
Metaverse Community Forum on Bullying and Harassment (2022)	Meta Community Forum: Results Analysis	Research report	<a href="https://cddrl.fsi.stanford.edu/publication/metaverse-community-forum-results-analysis">https://cddrl.fsi.stanford.edu/publication/metaverse-community-forum-results-analysis</a>
	Applying deliberative democracy: Insights from our partnership with Meta	Blog post	<a href="https://www.bi.team/blogs/applying-deliberative-democracy-insights-from-our-partnership-with-meta/">https://www.bi.team/blogs/applying-deliberative-democracy-insights-from-our-partnership-with-meta/</a>
Democratic Inputs to AI grant program (2023)	Democratic inputs to AI	Blog post	<a href="https://openai.com/blog/democratic-inputs-to-ai">https://openai.com/blog/democratic-inputs-to-ai</a>
	Democratic inputs to AI grant program: lessons learned and implementation plans	Blog post	<a href="https://openai.com/blog/democratic-inputs-to-ai-grant-program-update">https://openai.com/blog/democratic-inputs-to-ai-grant-program-update</a>
Collective Constitutional AI (2023)	Collective Constitutional AI: Aligning a Language Model with Public Input	Policy memo	<a href="https://www-cdn.anthropic.com/b43359be43cabdbe3a8ffd60ea8a68acf25cb22e/Anthropic_CollectiveConstitutionalAI.pdf?ref=openml.fyi">https://www-cdn.anthropic.com/b43359be43cabdbe3a8ffd60ea8a68acf25cb22e/Anthropic_CollectiveConstitutionalAI.pdf?ref=openml.fyi</a>
	Collective Constitutional AI: Aligning a Language Model with Public Input	News update	<a href="https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input">https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input</a>

## Appendix 2. Summary table for the four case studies.

Commissioning company   Case study	Commissioning context	Purpose and remit	Outcomes
<b>Google DeepMind, the RSA Citizens' Jury on Automated Decision-making (2018)</b>	Commissioned by Google DeepMind and designed and delivered by the Royal Society for Arts (RSA) as part of their Forum on Ethical AI. The timing of the jury followed a health data privacy controversy around Google DeepMind's training data. Two workshops with individuals from BAME backgrounds were convened in parallel.	Set by the RSA and an independent expert advisory group. The Citizens' Jury deliberated on the remit: "Under what conditions, if any, is it appropriate to use automated decision systems?"	The RSA's sought to engage influential companies and policymakers with the jury's recommendation.
<b>Meta, Metaverse Community Forum on Bullying and Harassment Deliberative Poll (2022)</b>	Commissioned by Meta, designed and delivered by the Stanford Deliberative Democracy Lab.	Set by Meta and its selected Advisory Board, the remit for deliberation was: "How to regulate bullying and harassment in virtual reality, particularly in the new private or 'members only' social VR spaces that are being created in the Metaverse?"	Followed by a second Deliberative Poll: Metaverse Community Forum on GenAI to test the reliability of Deliberative Polls adequate for the scale of global user deliberation.
<b>OpenAI, Democratic Inputs to AI grants program (2023)</b>	Commissioned by OpenAI, the company selected ten models of deliberation related to AI use and governance to receive grants to implement their designs.	Set by OpenAI, inspired by works of academics such as Aviv Ovadya and Helene Landemore. The problem posed for grant applicants was: "How to design a democratic process to govern AI?" Seven policy questions were specified for the applicants to develop their proposals in line with.	The grant program created an opportunity to assess the potential of the prototypes and inspired OpenAI's own research to set future directions.
<b>Anthropic, Collective Constitutional AI (2023)</b>	Commissioned by Anthropic and delivered by the Collective Intelligence Project (CIP) using the pol.is platform. The agenda was set by Anthropic to generate data on the similarities and differences between "in-house" and "public" values to govern AI.	Set by Anthropic to design a deliberative process identify public preferences on rules to govern the AI chatbot Claude and compare these against Anthropic's own constitution for Claude.	It is assumed to create a change to inspire companies and developers to continue exploring public inputs to AI.

# Acknowledgments

We are grateful for the contributions of Nicole Curato and Melisa Ross in reviewing this technical paper. We are also grateful for our interviewee and colleagues who provided their thoughts on our initial observations and helped us shape our thinking.

Nardine Alnemr worked at the Centre for Deliberative Democracy and Global Governance, University of Canberra while developing an earlier version of this Technical Paper. She acknowledges the Centre's immense support.

**Cite text as**

Canning Malkin and Nardine Alnemr (2024) Big Tech-driven Deliberative Projects. Global Citizens' Assembly Network (GloCAN) Technical Paper No. 5. Available at: <http://glocan.org/wp-content/uploads/2024/05/Technical-Paper-5-2024-Malkin-and-Alnemr.pdf>

**Funding declaration**

Research on the governance of citizens' assemblies was funded by the European Climate Foundation.

**Ethics declaration**

This research was approved by the University of Canberra's Human Ethics Committee (ID: 13354: Governance Review of the Global Assembly)