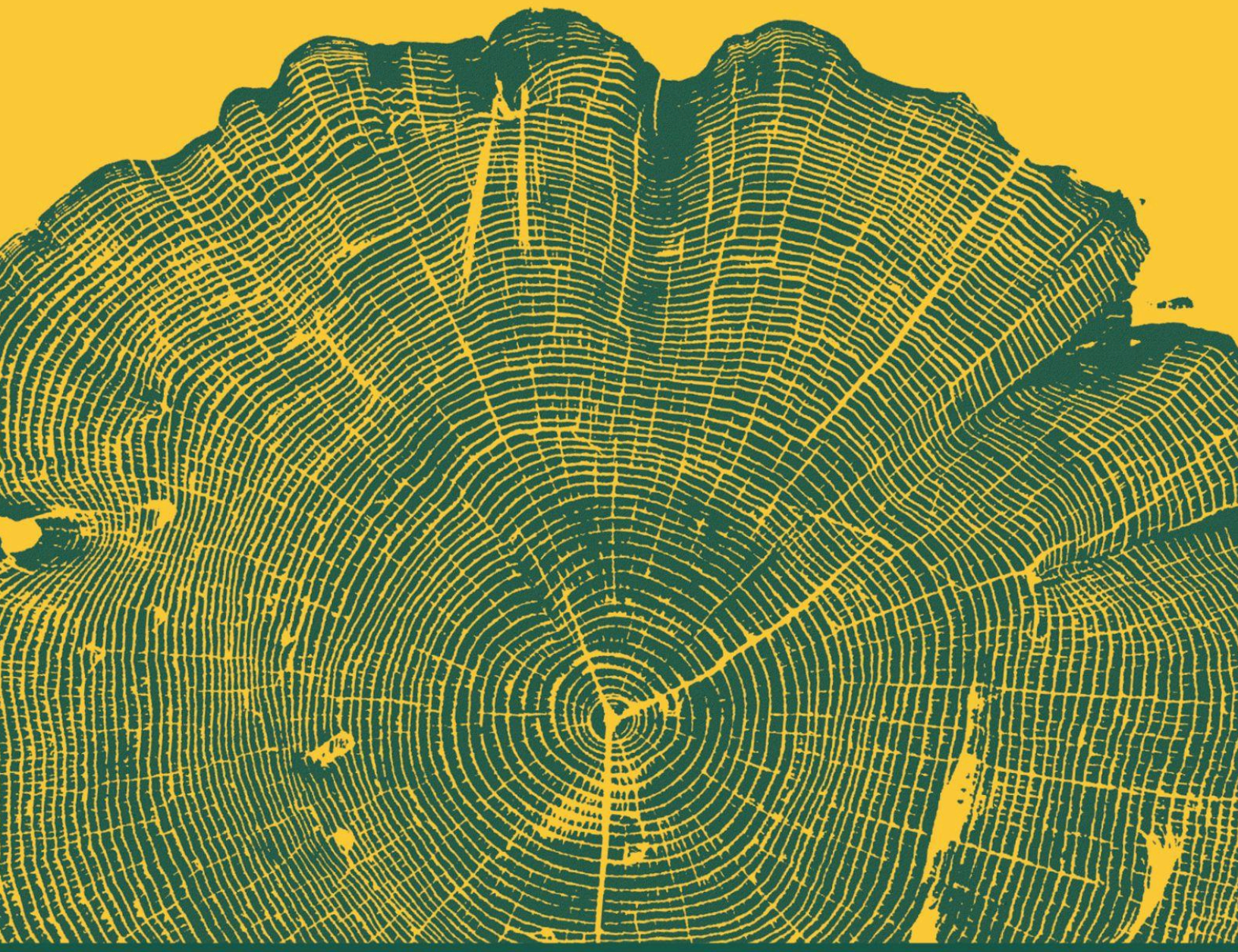

A ROADMAP TO DEMOCRATIC AI

March 2024

The Collective Intelligence Project

CONTENTS

The Collective Intelligence Project **3**
Introduction **4**
The Destination: What would democratic AI look like in 2030? 7
The Journey: What are our priorities for 2024? 8
Roadmap Guide: What can I do if I work in...? 9
The Roadmap **11**
Our Starting Point: What did we do last year? 12
Future Directions **13**
1. Advance the collective fine-tuning of generative models 13
2. Introduce collective input into more points within the AI lifecycle 16
3. Connect the open source and democracy communities 21
4. Expand public input processes geographically 23
5. Build AI-enabled tools for democratic governance 25
6. Experiment with institutional governance models 27
Conclusion **29**
Authors and Acknowledgements **31**
Appendix 32

The Collective Intelligence Project

The Collective Intelligence Project (CIP) is an organization creating better, more collectively-intelligent models of governing the transformative technologies that will shape society. Our partners have included OpenAI, Anthropic, Taiwan's Digital Ministry, the UK Frontier AI Task Force, the Creative Commons Foundation, and others.

We are a US-based nonprofit and are funded entirely by grant donations. This work has been generously supported by the Ford Foundation, the Omidyar Network, OneProject, the Survival and Flourishing Fund, and the Amaranth Foundation. For more on our operating model, see **https://cip.org**.

INTRODUCTION

For all the hype and investment, we're still in the early days of AI. We can still take the democratic path–and achieve the 2030 vision of collective stewardship and distributed benefits we've laid out below.

However, the democratic path is not the default path. Our social institutions—corporations, government, bureaucracy—are not currently equipped to take on this task. Democratic innovation is a public good: it is systematically under-provided without intervention. This is especially true for democratic innovation in governing AI. Practically, it's hard to develop better decision-making and distribution mechanisms that match the speed, focus, and concentration of resources driving the world's shiniest technology. It's hard to improve collective intelligence at the rate we're improving artificial intelligence. But that's what we need to do.

We founded the Collective Intelligence Project to find a new default path, and to build a better future. This work requires experimentation, commitment, resources, partnership, and coalition-building. We're grateful for the opportunity to work alongside, and learn from, inspiring colleagues who share our goal: to ensure that progress, participation, and safety don't have to trade off.

Why democratic AI? We think of democracy as more than deliberation, public input, or elections. At its core, democracy is a set of adaptive, accountable institutions that process and act on decentralized information, provide public goods, and safeguard people's freedom, wellbeing, and autonomy. When we say democratic AI, we mean an AI ecosystem that does the same, by default. **This document is our attempt to concretely describe what can be immediately done, built, researched, advocated for, and funded in 2024 in the AI ecosystem to achieve that goal.**

It's worth saying up front: **We do not think this document is exhaustive.** We don't discuss AI's impact on the nuts and bolts of existing, nation-state democracy. We don't cover the necessary role of stronger labor movements or a robust and expanded social safety net, nor do we discuss many ways we think AI could be used for direct public benefit, from healthcare to public services to education. We are an R&D lab at heart; our focus here reflects this.

Finally, this is a living document. We believe in collective intelligence; naturally, we also believe we're probably missing something that you know. **If you have an idea or a good example that we missed, if you vigorously disagree and are willing to walk us through your reasoning, or if you want to collaborate on the next steps below, please reach out to us at hi@cip.org.**

We're still early, but that doesn't mean we have much time. If you share this vision, we want to work with you, build with you, and support you.

LET'S DO THIS.

"AT ITS CORE, DEMOCRACY IS A SET OF ADAPTIVE, ACCOUNTABLE INSTITUTIONS THAT PROCESS AND ACT ON DECENTRALIZED INFORMATION, PROVIDE PUBLIC GOODS, AND SAFEGUARD PEOPLE'S FREEDOM, WELLBEING, AND AUTONOMY."

The Destination:

What would democratic AI look like in 2030?

We expect the world will look quite different in 2030. But here are three pillars we think are necessary for democratic AI, and provide especially fertile opportunities for experimentation, advocacy, and research.

This list is by no means exhaustive. Instead, in the spirit of democratic collaboration, we thought it was important to have a point of view and stake out a perspective where others may disagree.

1. **Our capacity for collective intelligence begins to keep pace with progress in artificial intelligence.** We use AI to improve our systems for deliberation, translation, facilitation, and preference elicitation; We expand this to improve interactions within and between institutions, as well as between institutions and individuals.
   Impact: We'll be much better at understanding and actualizing collective preferences.

2. **Important AI systems are governed by feedback loops of collective input.** We have built the infrastructure to inclusively gather, parse, and incorporate public input on complex questions surrounding AI. This includes responsively building AI systems to target real community needs, and enabling responsive opt-out of AI systems.
   Impact: Collective preferences directly inform any high-impact systems.

3. **High-impact sites of AI development are (re)structured to optimize for the collective interest.** The core material inputs to AI (data and compute) are governed non-monopolistically. Sites of development and deployment, whether open source movements, corporations, startups, or government agencies, are subject to checks and balances to mitigate against power centralization.
   Impact: We're not just gathering collective input, we've shifted incentives and built institutional capacity to actually compel action based on the public interest.

We also want there to be a variety of models that all sorts of people can use safely, opt-out guarantees for users, and strong data protection rules — not to mention many positive use-cases for AI. But these three achievements would massively change the political economy of AI, and they form a foundation for our vision.

The Journey:

What are our priorities for 2024?

This list starts off with extremely specific research advances we're building on, and then gets increasingly wide-angled. We believe both incremental advances and reaching for ambitious improvements in collective decision-making are equally important.

1. **Advance collective fine-tuning of frontier models.** In 2023, we prioritized our work on Alignment Assemblies, where we assembled people and communities to assess, deliberate over, and co-create AI models that reflected their values. This builds on work CIP and others have done successfully is a gateway to more democracy in AI.

2. **Identify other opportunities for ongoing public input in the AI lifecycle.** There are multiple opportunities in the development, deployment, and post-deployment of AI to incorporate ongoing public input. We will continue to develop better methods to assess what 'good' means, and figure out how to integrate the input within the technology itself in a straightforward and meaningful way.

3. **Connect the open source and democracy movements.** Efforts to increase access (open source) to AI, and efforts to increase participation (democracy) tend to attract different communities, but we believe that these two worlds – *hackers* and *participedians*, to simplify – need each other, especially when it comes to data governance, model fine-tuning, and risk assessment.

4. **Expand Alignment Assemblies to other parts of the world, and to other languages.** We must continue the push towards greater overlap between those affected by AI decisions and those making them.

5. **Build AI-enabled tools for democratic governance.** The future of democracy could be much better than the past. Building transformative technology into governance ensures that collective intelligence processes keep pace with AI and new possibilities for participation can be unlocked.

6. **Experiment with institutional governance models for AI development.** We're far from the best containers in which to build transformative technology. If we're going to achieve collective data governance, direct stakeholder input, and public accountability, we'll need to experiment with new development and funding models.

Roadmap Guide:

What can I do if I work in...

AI Research

- **Build participation technology.** Create AI-enabled collective intelligence tools and processes for better collective decision-making, including translation, moderation, facilitation, and preference aggregation. (see section 5)
- **Develop more processes for public input into AI systems.** Collective Constitutional AI is an example—innovate on collecting granular preferences and training models. (see section 1)
- **Identify new leverage points in the AI lifecycle for collective input.** (see section 2, especially 'Development' and 'Deployment')
- **Lead on collective governance of training data and improve the data supply chain,** including opt-out and transparency processes, self-determination for data laborers, etc.
- **Work with diverse audiences and communities to co-create models.** Engage people from different domain areas (democratic innovation, open source development, etc.) and geographies to apply these ideas in practice.

AI Development

- **Amplify internal teams already doing research on democratic AI research**. Our Collective Constitutional AI (CCAI) project with Anthropic and DeepMind's Habermas Machine are examples. Build internal leaderboards around success and breadth of public input.
- **Build out access to technology for developing AI-enabled collective intelligence tools and processes.** The future is wild, and we need to give people access to the best possible tools to improve our ability to coordinate. (see section 5)
- **Develop practical principles for embedding public input into internal organizational decision-making in a meaningful way.** (see section 2)
- **Explore collective input options for post-deployment monitoring and feedback.** This includes community oversight committees or councils. (see section 2, especially 'Post-Deployment')
- **Fund research into these topics.** OpenAI's 'Democratic Inputs to AI' grant scheme is one possible template for future work.
- **Include public inputs into internal evaluations and audits.** One example is our work with OpenAI (see section 2, especially 'Deployment' and 'Post-Deployment')
- **Look at alternative governance mechanisms** within your organization and supply chains. (see section 6)

Policy, Public Investment, and Regulation

- **Harness collective intelligence systems to assist with monitoring and evaluation.** This is especially true for socio-technical evaluations, societal 'red-lines', and monitoring societal impacts. (see section 2, especially 'Post-Deployment')
- **Develop more robust mechanisms to engage the public when creating standards and regulations.** This could enhance problem identification, solution ideation, and generate broader public buy-in. (see section 2)
- **Invest in public AI infrastructure.** Building expertise and resources outside of labs is crucial to ensuring broad public accountability; investing in applications of AI for the public good requires public sector AI infrastructure. This should include citizen and stakeholder engagement from the start, including in the allocation of resources (compute, data, etc.) and in public sector generative AI rollout decisions. (see section 2 and section 6)

Civil Society

- **Develop and establish new models for data governance.** This includes ideas like data sovereignty experiments or data cooperatives. (see sections section 2, especially 'Development', and section 6)
- **Support Alignment Assemblies internationally.** (see section 4)
- **Publish a leaderboard** that assesses how well AI companies incorporate the public interest, to support the directing of public contracts and broader support.
- **Support more equitable forms of data labor** for creating, labeling, and cleaning data. (see section 2 and section 6)
- **Shift the political economy of AI development from the bottom-up.** Explore economic models for public interest approaches to development, deployment, post-deployment, and funding e.g. cooperatives, crowd-funding, and more. (see section 6)

Open-Source

- **Actively connect with the democratic innovation space to create shared expertise,** recognizing that the democratization of access without governance rights is not enough to ensure the public interest. (see section 3)
- **Build AI-enabled tools for collective intelligence.** (see section 5)
- **Explore collective fine-tuning** on open source generative models. (see section 1)
- **Explore alternative public input mechanisms** to open source models. (see section 2, especially 'Development',)
- **Support Alignment Assemblies internationally.** We especially need technical expertise (see section 4).
- **Support experimentation with alternative governance infrastructures**, bringing in learnings from open source governance. (see section 6)

THE ROADMAP

Our Starting Point:

What did we do last year?

Before mapping the future in detail, it's worth going over how we've spent the past year. CIP's primary focus so far has been Alignment Assemblies (AAs). In the year we've been running these processes, we're proud to say we've worked with key partners from throughout government, industry, and civil society. We've also collaborated with and advised the UK AI Safety Institute on approaches to societal impact evaluation of frontier models, and on frontier AI impacts on democracy in the biggest election year in history.

We are incredibly proud of our work. **We've also constantly learned there is more to democracy than public input – which is why we've been so careful, experimental, and intentional about how we've structured each process.** At every step of the way, we've had to toggle between identifying new technological tools and new ways to structure processes so they can function as efficiently and effectively as possible.

Our work included:

Start Date	Partner	Result
March 2023	*Alignment Assembly:* **U.S. State Department Summit for Democracy**	Identified points of consensus between delegates to the Summit for Democracy on core questions in AI governance (more details).
August 2023	*Alignment Assembly:* **Taiwan Ministry of Digital Affairs (MODA)**	Formally identified public preferences for Taiwan's government policy on generative AI (more details).
October 2023	*Alignment Assembly:* **OpenAI**	Produced a usable, ranked list of risks most concerning to the US public when considering LLM impacts (more details).
October 2023	*Alignment Assembly*: **Anthropic**	Trained Anthropic's Claude model on a constitution written by a representative set of 1000 Americans (more details).
October 2023	*Alignment Assembly*: **Creative Commons** **We supported this experiment run by Shannon Hong, Kat Walsh, Timid Robot Zehta, and Nate Angell.*	Established six principles to govern how Creative Commons should respond to the use of CC licensed work in AI training (more details).
November 2023	*Collaboration:* **U.K. AI Safety Institute**	Built out program for societal impact evaluations, and frontier AI and democracy (more details).

Future Directions

I. ADVANCE THE COLLECTIVE FINE-TUNING OF GENERATIVE MODELS

OVERVIEW

Generative models are becoming more widely used, and model behavior may not always match the culture, language, or values of its users. Shaping model behavior to match these collective preferences is called "collective fine-tuning."

Our collaboration on 'Collective Constitutional AI' (CCAI) with Anthropic is one example of early work in this space. In this project, we gathered input from a representative sample of Americans to co-create a constitution to guide model behavior. We also quantified the benefits of this approach through quantitative and qualitative evaluations: the collectively created model behaved better than the control model on all nine dimensions of bias (e.g. gender, disability, race, socioeconomic status), while displaying equivalent levels of competence on a range of tasks. On March 4, Anthropic released the newest version of Claude, its foundation model, and it was based on the constitution we created.

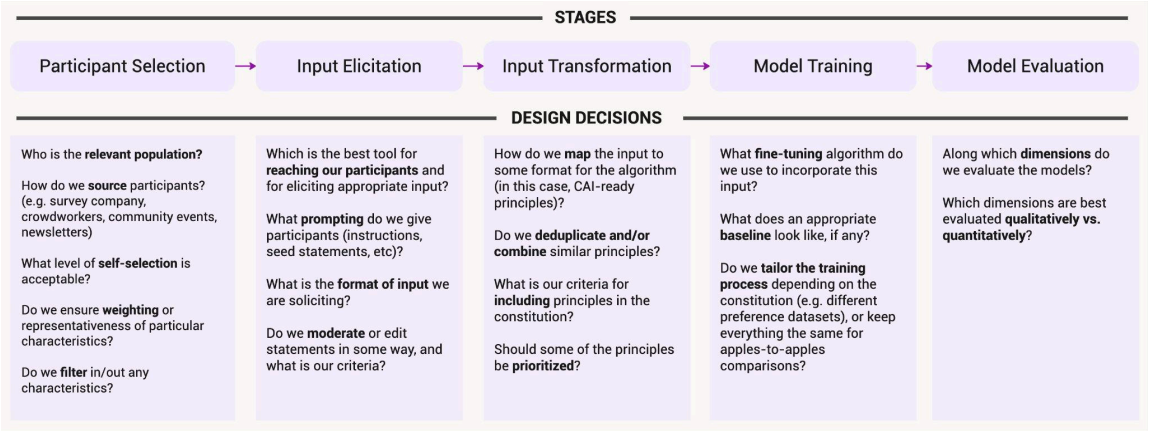

Figure 1: outline of our Collective Constitutional AI design decisions

Other AI labs could expand, scale, and implement this approach to their training. There is still a lot more work to be done on how to incorporate preferences, and we need more thoughtful tinkering and research, but this is a great first step. The open source community can also play a large role here, as transparency can enable trust in the results and encourages wider participation. We also think the opportunities for comparatively more rapid and dispersed experimentation and iteration could increase our rate of learning significantly.

PROMISING NEXT STEPS

Collectively fine-tune with different kinds of data. For example, how can we leverage existing datasets and pre-existing documents produced by communities as preference information?

Enable more groups of people to shape AI by making more robust and easy-to-run versions of the Constitutional AI algorithm. We probably could not have run the Constitutional AI algorithms without working with the original developers, which poses an issue for the ability for others to replicate and expand upon this work; we are now working on an open source version of this.

Open-source reward models and other methods of representing collective preferences.

Create more algorithms that can accommodate, represent, and/or bridge differing values. X's Community Notes is one of the most promising advances in collective intelligence from the past year, despite the contention around its corporate governance.

Create user interfaces to AI systems that make collective input easier or more effective. The Meaning Alignment Institute's work is an excellent example of this.

Train and release multiple collectively fine-tuned models with rigorous evaluation to establish how differences in technical and preference-elicitation approaches affect outcomes. Additionally, we should continue research to understand how value or preference input leads to changes in output. This will include A / B testing, visualizations of shifts, and deeper explainability research.

Reach out to people who would otherwise be unlikely to be able to participate in collective fine-tuning. This includes partnering with communities to fine-tune models for different populations, languages, and cultures.

2. INTRODUCE COLLECTIVE INPUT INTO MORE LEVERAGE POINTS WITHIN THE AI LIFECYCLE

OVERVIEW

Collective fine-tuning of models is a promising initial strategy for public input to AI, but there are many other leverage points throughout the AI model lifecycle. We need to identify and take advantage of those opportunities.

By collective input, we mean more than just collating preferences of users – AI systems are industrial processes that are fueled by data, labor, and compute, and the most effective ways to ensure accountability throughout the lifecycle is to develop granular knowledge of how these inputs and processes function in the real world.

We have split this lifecycle into development, deployment, and post-deployment stages, and have begun to identify avenues for experimentation.

- **Development:** the technical process by which deployable models are built. This includes the dataset construction, algorithmic design, base model training, etc.

- **Deployment:** anything else related to how the model is governed or released that does not directly relate to how the model is built. Two loose categories of work here that are potential levers for public input are "pre-release governance" (action taken before any kind of release) and "interface" (how people interact with the model).

- **Post-deployment:** Many important harms and impacts cannot fully be anticipated or mitigated before the model is released to the public, which makes responsive governance of the post-deployment stage very important and often overlooked

Future work will take two forms within each category: continuing to identify where the leverage points are and exploring what processes would work best for each. We have laid out promising approaches across the lifecycle below:

PROMISING NEXT STEPS

Development

Work on better conditions for the data laborers who shape our AI systems. There are many promising approaches to this that we support, including the support of unionizing/coordination efforts, platforming stories in the media, and providing legal assistance in legislative efforts; we think Alignment Assemblies could be a useful vector for driving awareness and action (see section 4).

Advocate for, and develop government investment programs for public AI. This includes public pipelines for data, compute, and talent, with democratic governance and accountability baked in from the start (initial work being done by NAIRR with OpenMined speaks to this). Major government decisions around the development/resourcing of AI options should involve public consultation in prioritization stages and publicly accessible accountability measures, with the necessary infrastructure to do so e.g. government-mandated processes and/or a new independent 'watchdog'.

Create more malleable and pluralistic datasets. One possibility is innovative data labeling techniques, such as the Stanford Jury Learning approach, to be deployed as a method of contextualizing AI models.

Build data cooperatives, data trusts,collectively developed data guidelines and other forms of community data governance. This is intended to ensure consensual and aligned (e.g. financially-aligned) training data inclusion. Both Spawning and Cohere's Aya are promising examples of new approaches to model building that reflect this goal.

Work on the open source development of a suite of different reward models to be used for reinforcement learning from human feedback. This will allow us to better understand methods of fine-tuning models and driving forward value and preference alignment.

Deployment

Enable public input into pre-release evaluations that determine whether an AI system gets released at all. This can include collective and public engagement in developing specific test metrics, societal "red lines" for model release, and establishing internal guidelines such as "Responsible Scaling Policies."

Make civic engagement a core function of governmental AI Safety institutes. AI safety institutes have recently been established in the UK, USA, Japan, and ongoing discussions of more. Engagement strategies are important establishing factors when institutes are founded, so the public can be involved in designing evaluations and informed understanding of models. (CIP collaborated with the UK AI Safety Institute, establishing their social evaluation and democracy team.)

Experimentation with more diverse and representative red teaming. This can include indexing on demographic representativeness, or ensuring more domain expertise such as medical professionals red-teaming a model for healthcare integration.

Post-deployment

Engage the public in developing metrics for better understanding the impact of AI on society, and in continuous monitoring. This follows up on the work we've done with the UK AI Safety Institute. Public evaluation and engagement need to be an ongoing process.

Develop federated incident reporting processes by e.g. mandating usage data sharing from labs, whistleblowing protections, and collective platforms for incident reporting such as the AI Incident Database). Construct 'evaluation juries' to determine proportionate action to be taken after high-profile incidents.

Exploring gamified systems and accessible user interfaces to more easily direct collective input into evaluations of generative AI systems. This is inspired by existing work on 'serious games' and could be used to enhance our ability to effectively engage the public on these issues. In particular, we believe this could support existing gaps in multimodal evaluations.

COLLECTIVE INTELLIGENCE INTERVENTIONS:

Below are diagrams of different pipelines and decision areas with a suite of existing and possible collective intelligence interventions mapped on. This is a draft attempt to make sense of the interlocking web of research projects and approaches. This includes work we have done (green), work others have done (gold), and promising work yet to be done (yellow).

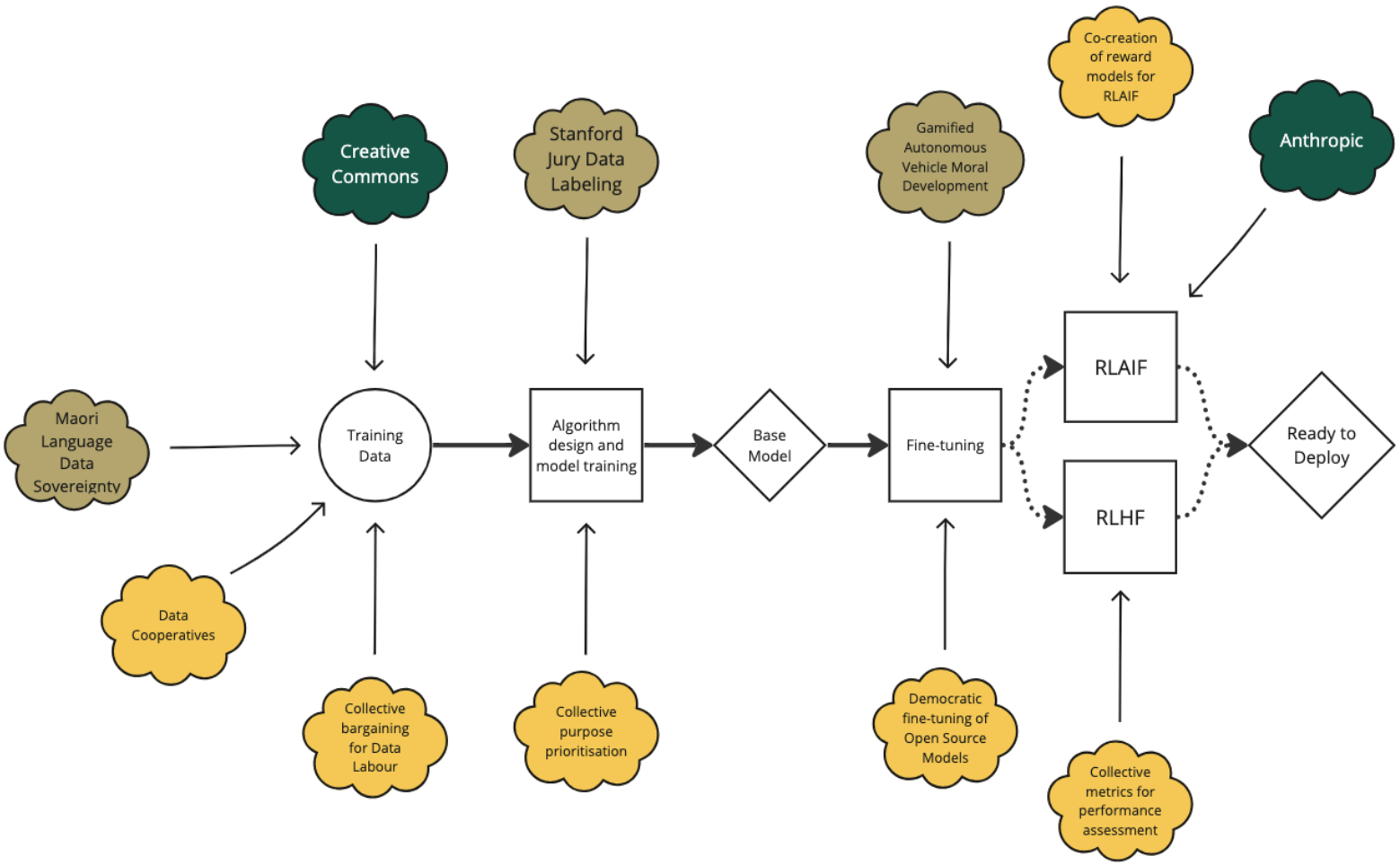

Figure 2: Generative AI pipeline with existing and possible collective intelligence interventions. Find more details in Appendix A.

AI Post-Deployment Ecosystem

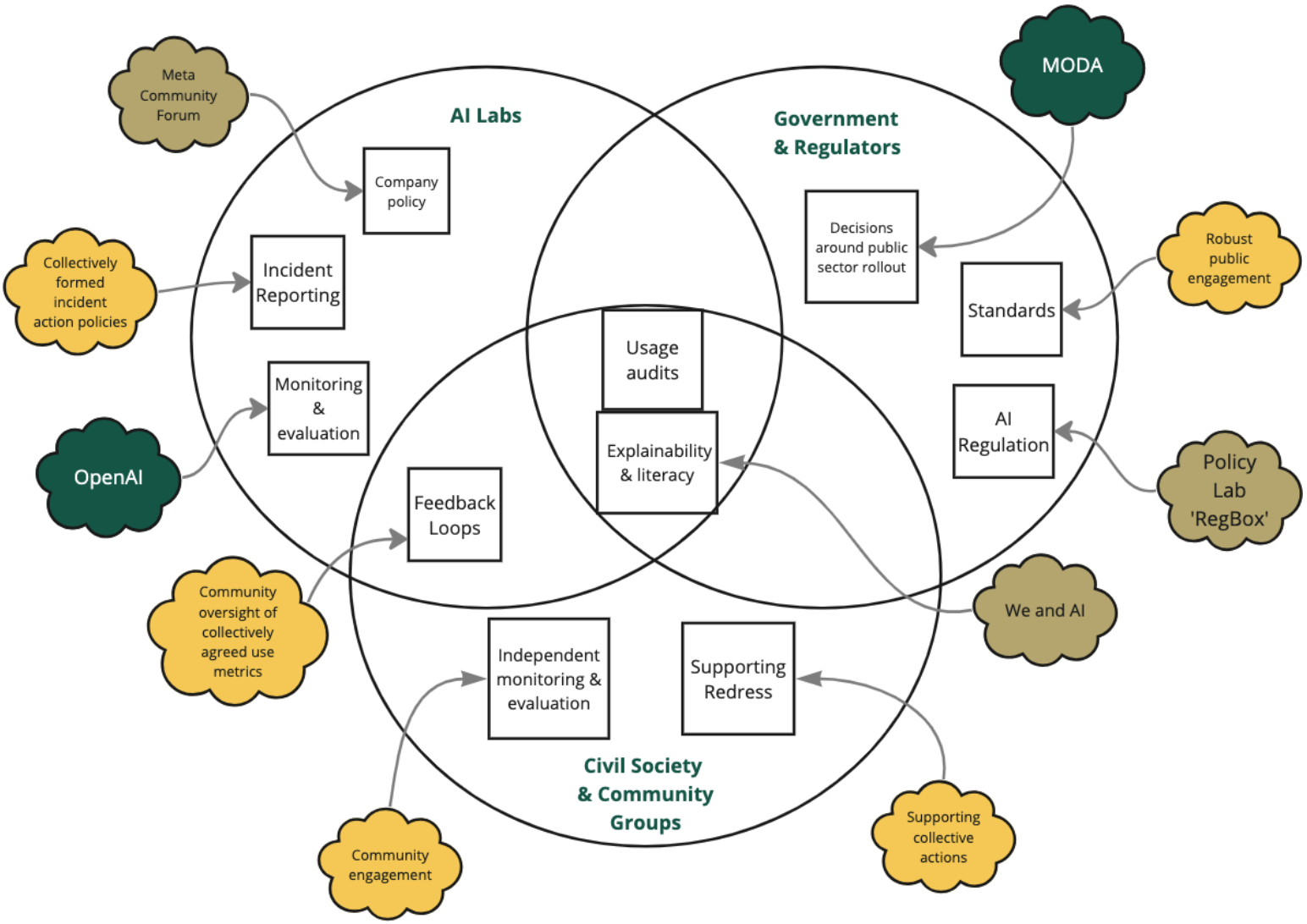

Figure 3: visualization exploring areas of responsibility for post-deployment governance, layered with existing and possible collective intelligence interventions. Find more details in Appendix C.

3. CONNECT THE OPEN SOURCE AND DEMOCRACY COMMUNITIES

OVERVIEW

The open source software movement has been a major force over the last three decades for increasing access to technology and making it more affordable. It is easy to assume that democratizing AI is as simple as investing in the open source AI movement, however, open source is not necessarily democratic, and "what people want" can be at odds with the direction of open source AI. Additionally, the direction of open source is ultimately dependent on the norms and rules of the programming communities that join together to make the software, choose the guardrails, and so on.

To democratize productive access to and governance of AI, we need to better connect existing open source initiatives with pathways to democratic oversight. We are excited about open source communities becoming closer with and more aligned with organizations working on democracy (e.g. public input delivery organizations like Stanford Deliberative Democracy Lab, Involve, Mission Publique, and others) and with communities impacted by, or seeking to use, LLMs (e.g. under-represented language communities).

More generally, we are excited about generating more connections between the open source and democratic innovation communities, to ensure that they are in the know about each other and can easily initiate collaboration opportunities.

This work should include work on open source LLM pre-training and fine-tuning and exploring different avenues for technology development and governance infrastructure. We are particularly interested in the distributed decision-making structures formed by open source collaboratives, such as those in the crypto (e.g. GitCoin for funding projects) or collaborative open source development (e.g. Gov4Git), and ways in which we could transpose this thinking to the domain of democratic AI.

This section is a little more open-ended than the previous ones, because we believe that these two communities could generate much better ideas, much more quickly, if they were in closer communication with each other.

PROMISING NEXT STEPS

Build open source tools for collective-fine tuning. This could include building a public input component into HuggingFace's recent work on Constitutional AI with Open LLM.

Create open source databases or platforms that allow for collaboration on new aspects of the AI pipeline beyond model access. This could include reward models, constitutional principles, fine-tuning data, and opt-out data.

Fund and run events, workshops, hackathons, and other activities to connect these spaces, developing shared expertise between the two. What happens when communities of hackers, and communities of process-oriented democratic innovators share a room together? We should also form specific collectives and organizations focused on facilitating this interdisciplinary collaboration and work.

Explore how historical open source work could translate to this new domain e.g. open source governance and resource allocation mechanisms. For example, Wikipedia-style editing models can be a treasure trove of collective preferences already; complex commons-based governance models that underpin many open source projects could be used to govern collectively fine-tuned models.

4. EXPAND ALIGNMENT ASSEMBLIES AND OTHER PUBLIC INPUT PROCESSES GEOGRAPHICALLY

OVERVIEW

Alignment Assembles were designed to bridge the gap between the people building AI systems and the people affected by them. This requires expansion to more places, communities, and languages. First, collective intelligence is improved by more and better information gathered from broader and more diverse sources. Second, a democratic approach to AI requires a frank accounting of existing inequalities in direction-setting access between the 'Global North' and the 'Global South', rural and urban areas, English and other languages, etc. Expanding geographically begins to address these issues.

This is not walking new ground: AAs will follow a rich tradition of participatory and democratizing practice in many countries around the world, from the origins of participatory budgeting in Brazil to pan-African initiatives in citizen engagement in policy-making, to Karya's AI-specific work in India.

With this expansion comes a range of predictable challenges: lack of expertise, poor access to infrastructure, language barriers, and lack of interest from more powerful entities, etc. To overcome these barriers, democratic AI practitioners should provide support to communities in the short-term and join the collective aim of co-creating shared resources with communities in the medium-long term, so they can empower themselves. In Appendix D, we've included a more detailed account of how to address these barriers, with examples of existing infrastructure to inspire work.

Our hope for 2024 is to see AAs in 5 more regions, which we would like to support with strategic advice and funding, if necessary. We would especially like to see AAs in countries with populations disproportionately affected by AI and those with large amounts of data labor.

Beyond straightforward expansion, we see multiple opportunities for innovation and exploration; further elaboration is in Appendix D.

PROMISING NEXT STEPS

Researchers should continue to build translation and education resources and interactive infrastructure for building collective intelligence on AI. This could look like wiki-style information databases and open source participation platforms like Pol.is and more.

Labs and companies should actively commit to a better supply chain for AI (including around data labor, compute, and other inputs). Model development is an industrial process that mirrors and replicates other forms of geographic extraction (through human labor, natural resource consumption, and more.). Governments–and savvy advocates–will play a key role in ensuring fair compensation and distribution of benefits, but concrete commitments by private entities could make things much easier.

Ensure that global fora for AI policy have genuinely international representation, rather than token inclusion. This should happen via experimentation, education, and advocacy targeting key areas like the UN, AI Safety Summits, Davos and others.

Governments should engage citizens in key decisions on AI development and use within their territory. CIP is proud of the work we've done with Taiwan and the U.K., and we're excited to work with other governments. We are also excited about the work that other organizations, such as The Ada Lovelace Institute, The Behavioural Insights Team, Sciencewise, Involve, and others have carried out around the world.

Build targeted AAs in collaboration with civil society organizations and government agencies that are already working with specifically affected communities.

5. BUILD AI-ENABLED TOOLS FOR DEMOCRATIC GOVERNANCE

OVERVIEW

AI has the potential to support incredible breakthroughs. While we are deeply concerned about the distribution of benefits, misalignment between the better angels of our nature and the reality of the default path, worst-case scenarios, and so on, we also believe that AI can be steered to enhance, transform, and improve democratic practice. It may be a problem, but it might also be a solution.

In-depth participation takes time, resources, and effort; complex decisions that happen at a national or international level are hard to adjudicate in this way – think of the difference between a lengthy, deliberative citizen's assembly and a low-fidelity but highly scalable national vote. Well-calibrated institutions are necessary (see section 6, and our broad agenda), but there are still technological barriers to meaningful, and productive, participation at scale. More complex decisions tend to have less granular participation, and vice versa.

AI systems could help mitigate the complexity vs. participation tradeoff by significantly augmenting reflective, deliberative, and executive capacity. As a start, LLMs could make translation – both in terms of actual languages, as well as terminology and concepts – much easier. We can also imagine broader forms of translation and facilitation — surfacing points of agreement and disagreement in deliberation, translating between viewpoints, identifying points of positive-sum convergence between two well-stated positions etc. For example, The AI Objectives Institute is exploring work on sentiment mapping, whilst Dembrane, Stanford, and others are working on facilitation.

This isn't a rejection of institutions and professionalism (see the next section) – instead, it's a vision where technology can improve the links between the realities of our daily lives and consequential decision-making. Recursive models (inspired by The Recursive Public and others) could support early participation in agenda-setting, shifting the balance of power early in decision-making processes when values are under contention, and support expert input at later stages that are focused on execution.

Beyond direct coordination, access to information – and making sense of the swarm of data that streams into our lives – is crucial for democracy. Current methods are often resource and time intensive – curating, personalizing, and walking through information tailored to audience needs requires enormous amounts of effort on cognitive, logistical, and formatting tasks. We hope AI can be steered to making this easier, higher quality, and more scalable.

We could also imagine more radical directions. For example, ideas like in-silico deliberation, delegating preference and value discussion to personal agents, and cooperative AI approaches to surfacing 'optimal outcomes' to satisfy collective preferences. We should, however, be incredibly cautious with these avenues: any form of delegation could lead to loss of agency and AI in its current form is often open to bias and 'hallucination'. However, it is hard to argue that existing systems are optimal for collective and individual agency, and more exploratory work on new directions is always welcome.

PROMISING NEXT STEPS

Experiment with cooperative models that pull out positive-sum outcomes for participants. Given knowledge of an information environment, build systems to identify best-case collective outcomes. Broadly, do more research on types of digital agents for representing needs, preferences, desires etc.

Build AI translation tools for instantaneous multilingual deliberation. Expand into translation capacity between viewpoints or other forms of inter-language translation.

Improving the digital or hybrid deliberation experience, particularly through analyzing large amounts of text-based data (as is being practiced by Civis). This might also be through better digital meeting room quality, visualizations, and more.

Use AI to support live learning, helping participants to drive deeper into issues or viewpoints as they emerge. Explorations of AI personal assistants and others lead in this direction.

Build better data analysis tools that can be used to track huge amounts of deliberative outputs to inform and connect discussions. For example, Talk to the City – our conversations matter, and we should be able to submit them for consideration beyond just whoever can text a politician.

Develop usable systems of real-time consensus mapping. Policy processes are feedback loops. It's worth exploring ways to build on The Recursive Public, which has harnessed AI already.

Expand work on how AI facilitators can support mass engagement in face-to-face, facilitated discussion. For example, Stanford's Online Deliberation Platform.

6. EXPERIMENT WITH INSTITUTIONAL GOVERNANCE MODELS

OVERVIEW

The above priorities focus on directly influencing AI models and tools. However, achieving the 2030 vision requires experimentation in the way we make decisions around developing frontier technologies as a whole.

In this final section, our main point is this: democratization requires collective infrastructure to ensure that the benefits of AI are broadly shared. Public input processes are necessary but not sufficient for democratization. A democratic AI ecosystem is one that is good for people, not just one that asks them questions at regular intervals. This system should ensure shared benefit from transformative AI, and that can both address risks and accelerate projects with the potential of broad positive impact.

Experimentation in this space is less straightforward than the others, but no less important. We welcome collaboration in trying out new containers for technology development. The history of transformative technology is one of creating new ways to develop tech — from the joint-stock corporation to the startup. There is much more innovation possible in the current era, building on models from open source projects, benefit corporations, focused research organizations (FROs), perpetual purpose trusts, cooperatives, decentralized autonomous organizations (DAOs), and more. Below are a set of possible directions to explore.

PROMISING NEXT STEPS

Build collective approaches to institutional decision-making.
Pre-deployment evaluations for frontier models are currently shallow; post-deployment evaluations barely exist. Internal corporate decision-making, legislation, and regulation are built on a foundation of evaluating impacts; collective input into these evaluations (whether through red-teaming, incident reporting, etc.) could be a beginning point for collective input into AI decision-making at the institutional level.

Build data cooperatives. In order to have technology that truly serves the common good, people must have control over their data as used for AI.

(P)redistribution. Labs like OpenAI have experimented with structures like a 'capped-profits' model; further commitments here may hold more weight than democratic input in terms of functionally improving people's lives. Democratic processes can be involved in determining how and by whom capital is distributed.

Createa permanent stakeholder council within AI labs. Building on existing experiments such as the Meta Oversight Board and Belgiums' permanent citizens' council, to create permanent bodies within AI development organizations to represent the collective interest, with membership determined by some combination of broad sampling, self-selection, and expertise.

Invest in public options for AI. DIfferent parts of the development and deployment pipeline could have public infrastructure components, from the collection of training data to the procurement or development of compute.

CONCLUSION

There is more to democracy than discussion and deliberation. There is certainly more to democracy than voting and elections. To meaningfully democratize AI, we will need to do a lot. We will need to shift the political economy of transformative technology towards the public interest, and away from short-term incentives and the extreme concentration of productive resources. We will need to engage with the whole of society.

We will need to imagine radically better worlds, and then work for them—not just for governance, but for the healthier, more creative, and more abundant future that these technologies could help support if driven towards the collective interest.

But we also need to start somewhere. This roadmap is our best current guess at what can be done, now, to lay the foundation for a democratic ecosystem for AI. We want to improve it, and we want to work on it with you: reach out at hi@cip.org for comments, collaboration, and critique.

AI has the potential to radically transform our societies for the better, but this is not a given. We must actively create more democratic methods of governing this technology and distributing its benefits if we want it to work for the collective good. As we said in the introduction to this document, *now is the time for ambitious investment and ambitious experimentation*.

LET'S DO THIS.

Authors and Acknowledgements

This report was written by the Collective Intelligence Project, a nonprofit research organization building collective models for the governance of transformative technology.

We are grateful to the following people for sharing their time and expertise with us as we constructed this roadmap. Some participants have chosen to remain anonymous.

Ana Colom, Public Participation and Research Lead, Ada Lovelace Institute
Audrey Tang, Digital Minister, Government of Taiwan
Austin Wu, Product Manager, Google
Brittany Smith, Head of UK Policy, OpenAI
Flora Pery-Knox-Gore, British Standards Institute
Jaron Lanier, OCTOPUS, Microsoft
Jennifer Ding, Senior Researcher, The Alan Turing Institute
Henry Farrell, Johns Hopkins University
Lama Ahmad, Policy Research, OpenAI
Lara Groves, Researcher, Ada Lovelace Institute
Liane Lovitt, Public Policy, Anthropic
Matilda Rhode, Digital Lead for AI + Cybersecurity, British Standards Institute
Maximilian Kroner Dale, Former Senior Advisor, Behavioural Insights Team
Shannon Hong, Fellow, Open Future
Tantum Collins, Director, US National Security Council
Vishal Maini, Mythos Ventures

Appendix

A: More detail on existing efforts in collective input for AI development.

Category	Work	Details
Our Work (with Affiliates)	Creative Commons	CIP affiliates ran two Polis-based assemblies at the Creative Commons annual conference in Mexico City and are expanding this to 500-1000 people in Feb 2024.
Our Work	Anthropic	Exploring use of representative samples in collectively developing constitutions to guide LLM behavior.
Others' Work	Maori Language Data Sovereignty	Through participatory initiatives that took place over 10 days in 2018 as part of the Te Hiku NLP project, the Māori community in New Zealand both recorded and annotated 300 hours of audio data of the Te Reo Māori language. The community established Māori Data Sovereignty Protocols to explicitly prevent corporate entities from owning the dataset. (p. 7)
Others' Work	Stanford Jury Learning approach	Introducing a supervised ML approach that chooses the set of labels informing a classifier based on the metaphor of a jury of data labellers, explicitly defining which people or groups, in what proportion, determine the classifier's prediction.

Others' Work	Serious game to crowdsource the public's views on moral decisions faced by autonomous vehicles	Researchers developed a website, Moral Machine, that used a 'serious game' with scenarios to crowdsource the public's views on moral decisions faced by autonomous vehicles. The aim was to generate a better understanding of the public's views about how autonomous vehicles should solve moral dilemmas, as well as to help raise awareness about this topic amongst the public. The platform was an effective large- scale data gathering exercise, collecting 40 million decisions in ten languages from people in 233 countries and territories.

B: More detail on existing efforts in collective input for AI Deployment.

Category	**Work**	**Details**
Our Work	Collective evaluations in partnership with OpenAI	We engaged a representative sample of 1000 members of the US public in a digital process to better understand public values and perspectives on the most salient risks and harms from AI. This fed into a roundtable with members of the public and staff from OpenAI to inform OpenAI's governance of their large language models.
Our Work	AI Safety Institute	We are planning at least one democratic input process to inform the work of the UK AI Safety Institute, scheduled to complete before Q2 this year.

C: More detail on existing efforts in collective input for AI post-deployment.

Category	Work	Details
Our work	Taiwan Ministry of Digital Affairs (moda)	We worked with moda to engage Taiwan's citizens in pursuing consensus around the opportunities and risks of frontier AI for Taiwan.
Our work	OpenAI	Our "Participatory Risk Prioritization" experiment with OpenAI produced outcomes for ongoing monitoring and evaluation e.g. monitoring for societal 'over-reliance'.
Others' Work	Meta Community Forums	Meta have collaborated with Stanford University and The Behavioural Insights team on two 'community forums'; engaging thousands of members of the public from around the world to explore key questions on their platforms and emerging technology.
Others' Work	We and AI	We and AI are a non-profit organization working to "encourage, enable, and empower critical thinking about AI". They aim to help more people make informed decisions about how they live with AI by engaging in public co-design, workshops, and more.
Others' Work	Policy Lab 'RegBox'	A toolkit enabling policymakers to convene stakeholders and work together to make decisions affecting regulation, using serious games.

D: More detail on addressing geographic inequities

BARRIER	SOLUTION
Expertise	Produce and compile guides on how to run these kinds of processes that can be shared and translated. **Examples:** RSA 'Democratising Decisions Around technology' toolkit We The Internet dialogue toolkits Global Assembly 'Community Assembly' toolkit Wellcome Responsive Dialogues Toolkit
Access to information	Collectively generated and updated resources on the latest developments in the space in order to support informed debate, hosted in an accessible way that can be updated and contextualized as is needed. **Examples:** Global Assembly Wiki
Access to infrastructure	Create/disseminate shared digital infrastructure for deliberation (e.g. open source platforms, community operated cloud service for hosting, etc.), information sharing, and for other logistical needs (e.g. low-fee financial infrastructure for transactions). **Examples:** Wikipedia (information infrastructure) Pol.is (deliberation) Open Collective (finance)
Language barriers	Ensure the availability of translation services, ideally from the local community. Write copyable resources and tools for translation into any language in an accessible format. Also, disseminate access to technology tools to support this.
Attention gap	Run shared processes across geographies to combine collective intelligence information, leveraging high-attention regions in bringing along the rest of the world.

CONTACT US

The Collective Intelligence Project

Website: cip.org
Email: hi@cip.org
Twitter: @collect_intel
LinkedIn: Collective Intelligence Project

The Collective Intelligence Project