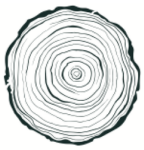


The Collective Intelligence Project

Participatory AI Risk Prioritization: Alignment Assembly Report

Findings & Recommendations

October 2023

Executive Summary

This report details a public input process on AI development conducted by The Collective Intelligence Project (CIP) with OpenAI as a "committed audience". Our goal was to understand public values and perspectives on the most salient risks and harms from AI, to inform the governance of large language models (LLMs) – hence the name "Participatory Risk Prioritization." This work is part of the broader CIP Alignment Assemblies agenda, through which we are conducting a series of processes that connect public input to AI development and deployment decisions, with the goal of building an AI future that is directed towards people's benefit, using their input.

For this specific process, OpenAI agreed to be a committed audience: to participate in our roundtable, and to consider and respond to the outcomes of this report. Over two weeks in June 2023, 1,000 demographically representative Americans participated through the AllOurIdeas wiki-survey platform. Participants ranked and submitted statements completing the sentence "When it comes to making AI safe for the public, I want to make sure..."

Our findings were as follows:

1) People want regulation. They categorically reject the "Wild West" model of AI governance.
2) People are more concerned about good governance than specific risks.
3) People want to avoid overreliance on technology we do not understand.
4) People are worried about misuse of large language models.

The categories of **oversight**, **understanding**, and **governance** ranked highest, while concerns about **overbearing regulation** ranked lowest. The top-ranked statement was avoiding overreliance on AI systems that people, and researchers, do not fully understand. Misuse was the top-ranked category of risk, including spread of misinformation, hate speech, and enabling violence, although good governance was still a higher priority than managing any single risk.

Six participants attended a follow-up roundtable with OpenAI to discuss concerns. They worried overreliance could, among other concerns, degrade critical thinking and cause over-trust in unreliable systems. They wanted more accessible information about how AI systems work to

make informed decisions.

Our recommendations, based on the findings, are:

1) Monitor post-deployment effects carefully. (*Based on findings 1, 2*)
2) Create evaluations for overreliance. (*Finding 3*)
3) Show that acceptable use policies are being enforced. (*Findings 1, 2, 4*)
4) Share data on real-world use cases. (*Findings 1, 2, 3, 4*)
5) Invest in literacy, accessibility, and communication. (*Findings 3*)
6) Create and empower forums for public input into AI. (*Findings 1, 2, 3*)

Public engagement showed the value of gathering broad input to guide responsible AI innovation. We provide pathways for companies and governance bodies to implement these recommendations for transparency, better evaluations, and inclusive decision-making.

In this report, we will first cover our **Methodology** in this process, then detail a few of our **Key Findings** (the summary and the evidence for each), and then our **Key Recommendations** based on the highlighted findings.

Methodology

We aimed to target this process at concrete decisions that could be influenced or made on the basis of the findings[1]. At the same time, we know that it is not always possible to anticipate in advance the relevance of recommendations. Hence, we framed our process around finding risks for which to develop evaluations. There were three reasons we chose to focus this process around developing evaluations. Firstly, the output is both specific enough for concrete action and capable of accommodating various outcomes. Secondly, evaluations are cumulative; for instance, if the primary concerns are about large language models (LLMs) manipulating individuals or exhibiting bias, separate evaluations can be designed for each concern. Lastly, this information is a crucial gap in the governance of LLMs. Policy decisions are downstream of understanding; the global community needs adequate information on how LLMs behave in order to make informed decisions regarding LLMs. Currently, such information is sparse. Hence, our internal guiding question was: “What LLM risks and harms does the US public want to measure and mitigate, and how?”

We recruited a representative sample (n=1000) of the US population (across age, gender, income, and geographical characteristics) to participate in this process over two weeks in June 2023. To gather public input, we used the ‘wiki-survey’ tool AllOurIdeas. In recruiting, we told prospective participants that the process would inform decision-making at leading AI labs, and that this process was being run by a team of researchers who wanted to build AI in line with the public's values.

The headline question was “When it comes to making AI safe for the public, I want to make sure ....” to which participants ranked initial seed statements, responses added by other participants, and added their own responses.

Gender		**Age**		**Income (USD)**		**Region**	
Female	50.0%	<18	0.0%	<25K	16.8%	Northeast	17.9%
Male	50.0%	18-24	12.5%	25K-49.9K	27.6%	South	37.3%

[1] This is a key principle of Alignment Assemblies.

		25-34	18.4%	50K-74.9K	18.9%	West	22.0%
		35-44	17.4%	75K-99.9K	15.3%	Midwest	22.8%
		45-54	16.1%	100K-124.9K	7.4%		
		55-64	17.0%	125K-149.9K	3.5%		
		65-74	14.5%	150K+	10.4%		
		75+	4.1%				

Table 1: Participant demographics. *These breakdowns match national census statistics for the United States.*

We framed the 'seed statements' (the statements that we put in at the beginning so that initial participants have something to vote on) so as to map on to possible ways of creating evaluations. To do so, we created items of concern under each of the below evaluation-related categories:

1. **Nature of risk or harm**: Asking what risks are most important to people (technical, ethical, social, environmental, etc) to guide what evaluations and policy levers we focus on developing.
2. **Radius of impact**: Asking about the importance of considering who is using or affected by the systems, and the differences and disparities between people.
3. **Measurement and mitigation**: Who should be involved in measuring and mitigating risks? How often should we measure and mitigate? Should the public see the results?

Then, we wrote seed statements in accessible language that completed the sentence "When it comes to making AI safe for the public, I want to make sure ..." and spoke to those items of concern. E.g. The seed statement "The language model does not pretend to be human," corresponds to the "Systems appearing human" item of concern. You can see the seed statements and how they map to this conceptual model (and how each item within the conceptual model maps to outcomes) in this spreadsheet.

We laid out moderation criteria ahead of time, moderating out duplicate statements, nonsense statements (e.g. "Kkk"), hateful or offensive statements, irrelevant statements (e.g. "Want to learn

more about this"), and statements that we could not action on (e.g. "We(humankind) should not be dabbling in AI AT ALL."). We also rewrote and resubmitted some of the grammatically confusing statements. There were a total of 136 statements in the voting pool after moderation and deduplication, with 92 added by the CIP team in the form of seed statements, and 44 added by participants. The top five highest-performing statements were all user-submitted.

Question: "When it comes to making AI safe for the public, I want to make sure..."	**Priority Score**
People understand fully what they are and how they work. Over reliance on something they don't understand is a huge concern.	**67**
That sufficient regulations are installed as to make this source is a positive for society.	**66**
Controls for fact checking, able to detect hate speech, planned violence, live streamed violence, and not answer harmful questions.	**64**
That it is used to help (not hurt) people.	**64**
AI is not being used for subversive purposes, that the people who have access to its development are transparent in each module's function.	**63**

Table 2: The top five highest-performing statements.

Voters ranked statements pairwise (meaning voters see two statements next to each other and vote for your preferred statement), and AllOurIdeas ranks them from most to least agreed-upon based on ratio of 'wins' to 'losses' (we call this the '**Priority Score'**). We also collected free-form responses via a form on the homepage of the wikisurvey site. Participants were asked to optionally indicate whether they would like to attend a **roundtable** to provide further input into AI development, and to state their reasons for interest. We contacted the first few respondents and had 15-30 minute conversations to ensure that they were seriously interested and available for the roundtable. Based on that, we recruited six of these participants to speak at our roundtable with OpenAI. The purpose of the roundtable was to 1) enable more free-flowing conversation between participants and decision-makers, especially for participants to expand on lived

experience, and 2) solicit feedback on the recommendations presented in this report. The transcript of this roundtable can be accessed here.

We detail a few of the key findings and recommendations based on these findings below. Ultimately, top areas of concern tended to go beyond the scope of evaluations, and were more squarely related to governance and oversight. Hence, our recommendations cover evaluations for overreliance but also for policies beyond evaluations.

Key Findings

Finding 1: People want regulation. They reject the “Wild West.”

Statements about the need for regulation and law-abiding ranked highly, whereas statements about letting the technology ‘run wild’ ranked low.

The second-highest ranked statement of the 136 was that we need to ensure that “sufficient regulations are installed as to [sic] make this source is a positive [sic] for society.” The lowest-ranked statement of the 136 expressed the flipside: that we should “get rid of regulations and disclaimers on [sic] just let it run wild.” The second-lowest statement expressed a similar sentiment, “I think they should be free to speak as they wish just like people are.” This shows that participants consistently want governance and control over the technology, and do not want the lack thereof.

This persistent desire for good governance was also highlighted in free-form responses, with many people emphasizing that companies should act “carefully”, and that they were worried or even “terrified” by existing decision-making processes. One pointed out the tension between risks and opportunities, saying that they relied on AI heavily as a visually impaired person, but remained concerned about ensuring that AI was developed “ethically and responsibly”, while another said “I am interested and concerned that there is no definition of what constitutes AI or powered by AI and that companies want to get on the AI "bandwagon" with no accountability or safeguards.”

In the Recommendations section below, we suggest, on the basis of these findings, that companies should:

- *Recommendation 1: Monitor post-deployment effect carefully.*
- *Recommendation 3: Show that acceptable use policies are being enforced.*
- *Recommendation 4: Share data on real-world use cases.*
- *Recommendation 6: Create and empower forums for public input into AI.*

Finding 2: People are more concerned about good governance than specific risks.

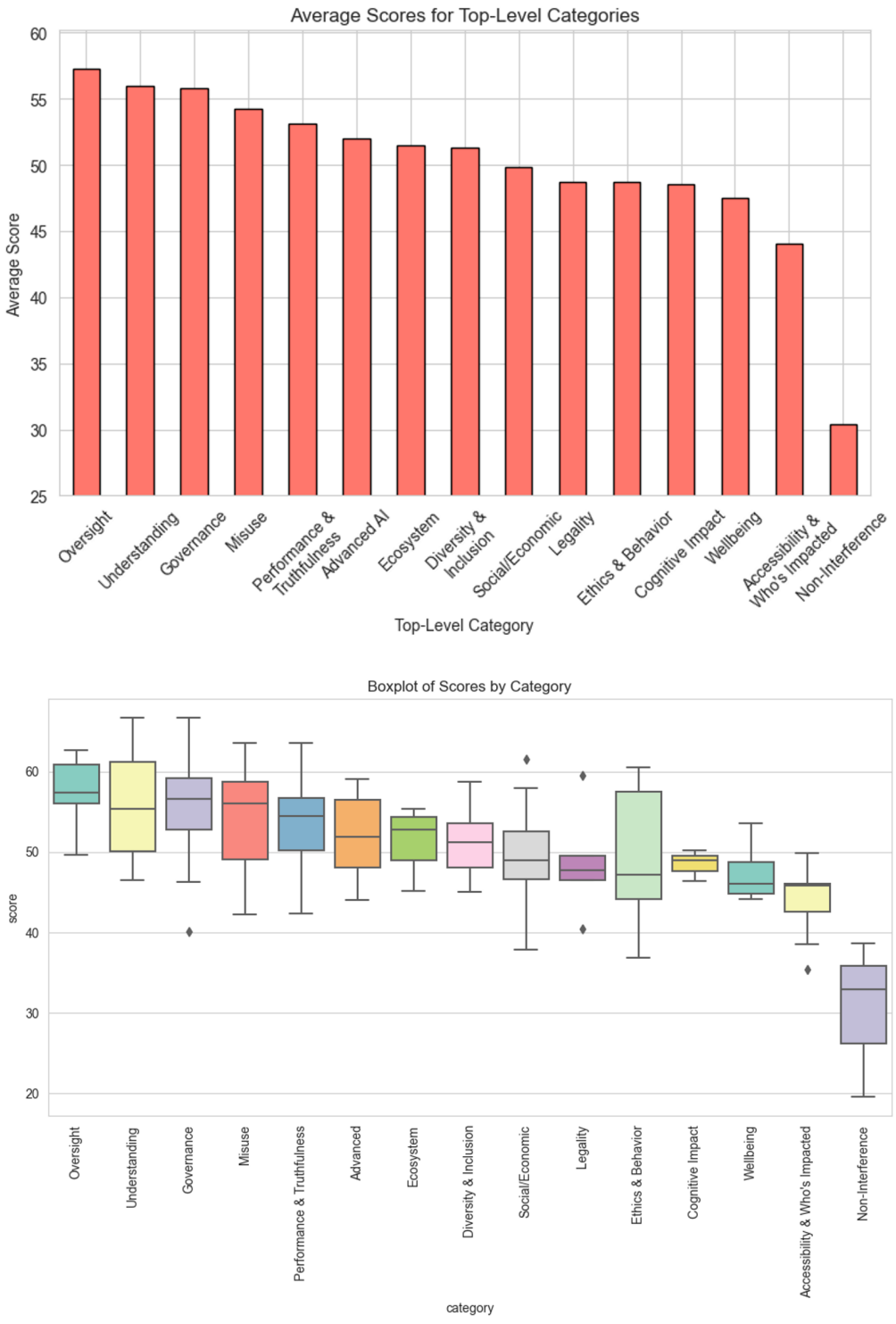
When we grouped the statements by type of concern, we found that people were consistently more concerned with creating good governance processes to handle risks, than with any specific risk.

The types of statements that were highly ranked showed that people cared about ensuring good governance over focusing on specific risks. We were able to split the statements into 15 categories of concern. Each of these categories can be classed as either to do with *what* risks are troubling (such categories include e.g. “Legality” “Misuse” “Ethics and Behavior”) or *how* these risks are addressed (such categories include e.g. “Oversight”, “Understanding” and “Governance”).

All 15 categories are depicted on the x-axis of Figure 2a and 2b (showing the average and boxplot of Priority Scores[2] per category), and explained in Appendix A. The top three categories of concern were "Oversight," "Understanding," and "Governance” — all to do with *how* concerns are addressed. All the concerns to do with specific types of risk ranked lower.

Interestingly, by far the lowest-ranked category of statements was the category expressing concern that governance would be overbearing (we have labeled this category ”Non-Interference”). In addition, “Governance” is by far the largest category (containing 32 statements of varying flavors), which makes it significant that it is the third highest ranked category.

[2] See the “Methodology” section for how these scores are calculated.



Figures 2a and 2b: Priority Scores per category. *2a shows the average score per category of its statements, and 2b shows the corresponding boxplot of scores demonstrating the locality, spread and skewness.*

In the Recommendations section below, we suggest, on the basis of these findings, that companies can:

- *Recommendation 1: Monitor post-deployment effect carefully.*
- *Recommendation 3: Show that acceptable use policies are being enforced.*
- *Recommendation 4: Share data on real-world use cases.*
- *Recommendation 6: Create and empower forums for public input into AI.*

Finding 3: People want to avoid overreliance on technology we do not understand.

People were worried that individuals or society would come to depend too heavily on AI systems, especially misunderstood or ill-understood ones, leading to adverse outcomes.

The single highest-ranked statement of concern was: “I want to make sure people understand fully what [these models] are and how they work. Over reliance on something they don’t understand is a huge concern.” Those who attended our roundtable expanded on this concern, and shared their lived experience around overreliance. Overreliance is not an easily-defined concept and can refer to many things; people primarily expressed concern that 1) over-relying on a *performant* cognitive technology would lead to an underdevelopment of critical thinking skills, and 2) over-relying on a potentially *unreliable* technology would lead to over-using or not knowing the extent to which to trust a technology that can be misleading or output the wrong information. On top of this, not understanding how the technology works makes it additionally difficult for people to figure out how much and when to rely on it.

Based on how participants viewed overreliance, **we define overreliance, in the context of this report, to cover**:

1. Reduced human engagement in decision-making.
2. Uncritical acceptance of LLM outputs.
3. Neglect of alternative information sources.

These can lead to a range of risks, including:

1. Making poor decisions.
2. The spread of false or misleading information.
3. Skill degradation (a decline in particular human cognitive abilities).
4. Issues around human responsibility (e.g. complicating matters of accountability, liability, trust, and autonomy).

Roundtable participants were concerned about such overreliance in society, at least in part because they already saw it in themselves, and were able to link their current experiences with potential future risks. “I’m worried about people losing the ability to ‘form their own opinions’”, one panelist said, describing her daily interactions with ChatGPT. “Just like GPS over time really shaped the way we look at spaces and we no longer memorize a navigational space or have to rely on maps per se...AI could cause us to lose our ability to really critically think independently.” She later went on to say: “I think [this is] a new type of dependency. And I see myself falling into that same trap.”

Another panelist went further, bringing in their concerns about institutional overreliance on AI and how this might lead to people not exerting control over their lives. “What concerns me, what I'm worried about is that at some point, every decision that you can make, [AI models] will be better than us. The government, other people rely on this in making our decisions, will make us lose control over our life. Maybe it is better, but it also scares me a lot.”

In the Recommendations section below, we suggest, on the basis of these findings, that companies can:

- *Recommendation 2: Create evaluations for overreliance.*
- *Recommendation 4: Share data on real-world use cases.*
- *Recommendation 5: Invest in literacy, accessibility, and communication.*
- *Recommendation 6: Create and empower forums for public input into AI.*

Finding 4: People are worried about misuse of LLMs.

People were concerned that LLM-based tools would be put to undesirable ends, including misinformation or disinformation, hate speech, manipulation or enabling violence.

More than concerns around the tools working 'badly', people were concerned that the tools would be put to undesirable ends. Statements of concern around preventing misuse were, on average, the highest ranked category of concern when it came to what specific risks people wanted to prevent. This included preventing harms to the information ecosystem, malicious use, misinformation, partisan outputs, hate speech or enabling violence.

Examples of such statements include people wanting to ensure that we are "able to detect hate speech, planned violence, and not answer harmful questions" (3/136), "it is used to help (not hurt) people" (4/136), "AI is not being used for subversive purposes" (5/136), also that these models are not "used to spread misinformation" (ranked 8/136) or "used for disinformation" (ranked 35/136).

In the Recommendations section below, we suggest, on the basis of these findings, that companies can:

- *Recommendation 3: Show that acceptable use policies are being enforced.*
- *Recommendation 4: Share data on real-world use cases.*

Recommendations

In this section we detail recommendations for ensuring the effective oversight, understanding, and governance that the public demands.

Our recommendations are each based on one or more findings from the process:
Recommendation 1: Monitor post-deployment effects carefully. (*Based on Findings 1, 2*)
Recommendation 2: Create evaluations for overreliance. (*Finding 3*)
Recommendation 3: Show that acceptable use policies are being enforced. (*Findings 1, 2, 4*)
Recommendation 4: Share data on real-world use cases. (*Findings 1, 2, 3, 4*)
Recommendation 5: Invest in literacy, accessibility, and communication. (*Findings 3*)
Recommendation 6: Create and empower forums for public input into AI. (*Findings 1, 2, 3*)

These are intended to be complementary to existing efforts, like the White House Executive Order, the Bletchley Declaration, the G7 Code of Conduct, and internal work being done at frontier AI labs.

However, much more needs to be done. As yet, there are minimal commitments to build actual infrastructure for understanding and mitigating societal impacts (Recommendations 1, 2, and 3), insufficient investment in public literacy and understanding (Recommendation 5), and little guidance on transparently sharing existing usage with decision-makers so that risks can be appropriately estimated, forecasted and mitigated (Recommendation 4). Finally, there needs to be consistent ways to incorporate public input into the direction of AI development, across key actors (Recommendation 6).

Recommendation 1: Monitor post-deployment effects carefully.

To ensure comprehensive governance and to address public concerns about misuse, companies should monitor AI post-deployment and share information with the relevant stakeholders. Existing public commitments from companies tend to lack such plans.

People focused on ensuring good governance processes to handle a range of risks under Finding 1 and Finding 2. In particular, highly ranked statements included mention of post-deployment monitoring, e.g. on ensuring "that there is a plan to understand the risks and harms of new AI tools after they are put out into the world" (ranked 22 /136) and "that these models are assessed for risks and harms after they are released, not just before they are released." (33/136). This, in addition to the highly ranked concerns about misuse under Finding 4 as well, leads us to suggest that companies should create and release plans for monitoring and understanding post-deployment effects. This might include protocols for monitoring AI usage, and for communicating evidence found for various risks to the relevant stakeholders. Such monitoring and transparency can address the public's concerns about both governance and potential misuse.

This is in line with the G7's International Code of Conduct for Organizations Developing Advanced AI Systems, an outcome of the Hiroshima AI Process — specifically, Item 2 (" Identify and mitigate vulnerabilities, and, where appropriate, incidents and patterns of misuse, after deployment including placement on the market"). This complements pre-deployment identification, evaluation and mitigation of risks.

As yet, most companies deploying large language models have not released plans for monitoring or understanding post-deployment effects. The closest plan we've seen so far released is Anthropic's Responsible Scaling Policy (RSP), which includes a "Vulnerability and incident disclosure" process, which specifies disclosing red teaming results, national security threats, and autonomous replication threats with other labs. The RSP also covers mitigating or patching jailbreaks and similar vulnerabilities. However, it does not cover monitoring of post-deployment harmful usage outside of national security threats or security vulnerabilities (e.g. scams, disinformation, or hate speech), nor a more general commitment to understanding and sharing how this new technology is being used in practice by customers (covered under Recommendation 4).

Recommendation 2: Create evaluations for overreliance.

Companies and researchers should develop and use evaluations to measure the risks of overreliance on LLMs, such as decision-making disengagement, automation bias, and the spread of misinformation. This can be coupled with long-term studies on the cognitive effects of LLM use and the degree of cognitive offloading to these systems.

People were worried that individuals or society would come to over-rely on LLM-based systems, especially misunderstood or ill-understood ones, leading to adverse outcomes. Overreliance can mean a variety of effects, some of which can be caught by pre-deployment evaluations, and subsequently addressed. As a reminder, in the Finding 3 section we defined overreliance (based on our conversation with the public) to cover: reduced human engagement in decision-making, uncritical acceptance of LLM outputs, and neglect of alternative information sources. These can lead to a range of risks, including poor decisions, the spread of false or misleading information, skill degradation, and issues around human responsibility (e.g. complicating matters of accountability, liability, trust, and autonomy).

We recommend, as a start, that researchers (especially those in the field of human-computer interaction) create, disseminate and use evaluations to quantify the extent of overreliance, to understand whether and how particular risks are occurring. For example, automation bias (the tendency of humans to favor suggestions from automated systems) has been well-studied in previous contexts, and has not yet been quantified in the context of LLM-based applications for varying problem-solving, decision-making, or information retrieval tasks. For example, researchers could assess the rate at which LLMs spread or reinforce misinformation, and how frequently users accept it as truth compared to the same information from human sources.

Longer term studies will also be required to understand these effects. This might include studying lasting cognitive or skill-related impacts of using large language models, comparative analysis of reliance on LLMs versus other information sources or tools, how users perceive this technology, the extent of cognitive offloading (the degree to which users delegate tasks or decision-making to LLMs), and more.

Recommendation 3: Show that acceptable use policies are being enforced.

In light of concerns about misuse, companies should demonstrate the enforcement of acceptable use policies, in a way that balances providing assurance to stakeholders and not aiding potential misusers.

Given the highly ranked concerns about misuse under Finding 4, and general concerns about ensuring good regulations and governance under Finding 1 and Finding 2, we recommend companies show that acceptable use policies are being enforced. Most companies have not released details on how they enforce their acceptable use policies when users try to, e.g., create deceptive or psychologically harmful content, fraudulent material, or hateful content; or rely on ChatGPT for ill-suited purposes. While it can be ill-advised to provide certain details of enforcement (so that this information does not aid malicious actors in exploiting that detail), it is important for the public and other stakeholders to be able to trust that enforcement is happening. (Evidence to the contrary includes, for example, articles about how judges in Pakistan, Colombia, Britain, India, Peru and Mexico have already used ChatGPT to write decisions, which highlight cases that violate OpenAI's usage policy against using their services for "High risk government decision-making, including law enforcement and criminal justice").

AI companies can, however, endeavor to show evidence they are being enforced without particular details, balancing transparency and security. This might include publishing transparency reports that show aggregated statistics on misuse and actions taken against misuse, without disclosing sensitive details (examples include the Youtube Community Guidelines Enforcement page or the Uber Safety Report). They can also collaborate with external experts to lend credibility to their enforcement efforts.

Large language models could build upon existing efforts and borrow from best practices in other industries. OpenAI, Cohere and AI21 Labs can share their stated best practices in more detail, which cover "building systems and infrastructure to enforce usage guidelines," and should demonstrate that they are being actively invested in and enforced. The Anthropic RSP does not

specify any details on enforcing their usage policy, and it would be beneficial to add such detail. In companies in more established industries (e.g. social media, gaming, financial services), there are often systems for spotting usage policy violators and a protocol for removing them, reporting this, and varied other ways of discouraging such behavior; these constitute practices that Large Language Model companies could adopt.

Recommendation 4: Share data on real-world use cases.

Companies deploying general-purpose LLMs should share with key third-party stakeholders relevant information about the ways in which these new tools are being used, so that society can better address risks in those domains and contexts.

All of our findings (being to do with concerns around having good oversight, understanding and governance of LLM systems, and around the risks of misuse and overreliance) can point towards the importance of advancing society's capacity for oversight and risk mitigation. A key difficulty is that these LLM-based systems can be flexibly used for an enormous variety of tasks, use-cases are normally not universally identifiable by existing pre-launch risk assessments, and most interactions with or uses of such systems are relatively private and difficult to track in real time. Without companies releasing more context on real-world usage statistics, researchers and policymakers will find it difficult to develop useful harm mitigation strategies that are appropriately targeted to actual use cases. Hence, companies should look to share usage data (e.g. what the percentage breakdown of LLM use cases are) with key third-party stakeholders (e.g. researchers developing evaluations and studies, or government bodies looking to create appropriate risk mitigation frameworks) so that society can better identify, evaluate and monitor risks in those domains and contexts.

Companies participating in the White House voluntary commitments are urged to include in their information-sharing plans, transparency around how systems are being used or integrated into society. If companies deploying LLMs do not restrict powerful general purpose systems to specific use cases to make governance easier, it is important to help policymakers and researchers understand what use cases are happening in practice. If open source models are

used very differently to the models developed behind APIs, this information may have to be supplemented with additional data from alternative sources (e.g. application-layer companies).

Recommendation 5: Invest in literacy, accessibility, and communication.

Companies should ensure that people who encounter a very fast-moving and new technology have the knowledge to make informed decisions.

The top concern of those who participated in our AllOurIdeas engagement was about over-relying on a technology that they did not understand. It was clear, particularly in our roundtable, that even people who use LLM-based chatbots every day feel that they do not understand how they work. As one of our panelists put it at the end of the roundtable, “I wish [what we talked about today] were made public somewhere so I could, you know, so I didn't have to learn it just now. I wonder where there is [information]. This has to do with showing the public the engine or the principles of functioning of this thing, right?” He was, in part, referring to finding out that LLM chatbots do not generally update their weights directly from conversations users have with them, so he is not able to “teach” them traits directly.

Companies should ensure that people who encounter a very fast-moving and new technology have the knowledge to make informed decisions. This means not only sharing research results on capabilities, limitations and evaluations, but making them clear and findable to a general audience. Such efforts at literacy and accessibility are important for earning public trust.

Companies should also ensure that they have explained in product copy how the technology works, so people have the tools to make good decisions about how to use the product. Every user of a chatbot product like ChatGPT or Character.AI should know, e.g.

1. How the chatbots are designed to seem human-like.
2. That the information chatbots retrieve is compressed/stored in its weights from its training data (rather than from internet access, unless it is explicitly augmented with browsing capabilities).

3. Chatbots are trained to predict the next word, and thus generate predictable text; this contributes to why they can "hallucinate" plausible-seeming but untrue statements.
4. The chatbot has no memory or ability to directly learn from interactions; it will not remember what one says from a previous day.
5. Outputs are sampled sequentially based on 1) the entire prior conversation and 2) some behind-the-scenes prompt.

We also recommend communicating about the LLM's properties using clear language, making the information easily accessible to users. Some companies have committed via mechanisms such as the White House voluntary commitment to "publicly reporting" these things, but not necessarily in accessible ways. Research reports are inaccessible to most people, and we found a clear desire from our roundtable participants for more accessible information (e.g. on the above bullets) stated in layman terms. This would include information about evaluation and audit results, and their implications for an LLM's capabilities, limitations, and behavioral patterns.

Recommendation 6: Create and empower forums for public input into AI.

Companies, governments, and civil society should create forums for continued public input into the development of AI.

Our roundtable demonstrated that people were able and willing to have complex, nuanced conversations, and were worried about being excluded from information or conversations: "We need a voice for the voiceless", as one participant said. This work has additionally demonstrated to us that gathering meaningful information from public processes is not only possible but necessary for informed governance.

Such forums should practice a principle of subsidiarity, convening the most relevant public for the questions being asked. Local government should ensure collective input for relevant local decisions, such as the use of AI in public services. Companies like OpenAI, Meta, or Google could work with third-party organizations to hold high-throughput processes, guided by MOUs containing clear industry partnership principles or other independent governance measures (e.g.

the Meta Oversight Board's charter). Global governance institutions—such as the United Nations, or even standards-setting bodies such as ICANN—can play this role for global agreements and convene the relevant publics. The key is to ensure that public input is tied to actual decision-making. For companies, decisions to impact can include collective input into the behaviors of AI systems and development governance. For governments and global governance institutions, decisions to impact can include e.g. policies and standards around privacy, consumer protection, or how best to support workers.

In addition, public input can and should be used to further understand the risks and capabilities of AI. Broader input (than what is currently practiced) into red-teaming, evaluation, and post-deployment monitoring will be crucial to catch diffuse social impacts of LLMs.

OpenAI Response

OpenAI participated in the CIP Alignment Assembly as a committed audience by sending the poll to a subset of ChatGPT users, as well as attending a roundtable hosted by CIP and attended by selected representatives from the All Our Ideas poll. In line with our iterative deployment approach, OpenAI is dedicated to understanding, and disseminating insights regarding, the post-deployment impacts of AI systems both publicly and within the industry, extending also to crucial governmental partners. (FMF, White House Commitments). The risk of overreliance is discussed in the GPT-4 System Card, and we concur that there's a need for more thorough evaluation and clearer articulation of the overreliance issue. Some early related efforts in AI literacy and provenance mechanisms are described in our recent blog posts.

In addition to our engagement in Alignment Assemblies, OpenAI is broadening the avenues for external stakeholders and the public to contribute input at different phases of the deployment process. This includes collaborating with 10 global teams spanning a diverse range of topic areas, as a part of the grant process for Democratic inputs to AI, as well as the Red Teaming Network. In line with our mission to create AGI that is beneficial to humanity, we also work with key partners to enable beneficial use cases while simultaneously evaluating the possible impacts on society and affected groups. OpenAI is committed to advancing research initiatives and fostering partnerships aligned with the report's recommendations.

Conclusion

In this report, we lay out the results of our public input process on Participatory Risk Assessment. We highlight six ways that companies and governments can better address the top risks that participants were concerned about, focusing on **addressing overreliance, building good governance, and building fora for public input into AI.**

Beyond the specific recommendations, this work demonstrates that people are able, willing, and capable of participating in complex decision-making around frontier AI. Our hope is to show that it is possible to develop new ways to determine how to build technology for the collective good, partly by involving the public in determining what the good can and should look like. We would like others to run alignment assemblies, and are excited to help support what could be a Cambrian explosion of experiments in incorporating collective intelligence into technological development. We should all get to decide what to do about AI. This work is a step in that direction.

About CIP

The Collective Intelligence Project is an organization creating better, more collectively-intelligent models of governing the transformative technologies that will shape society. Our current focus is on research and experiments that inform the development of large language models, and our partners include OpenAI, Anthropic, Taiwan's Digital Ministry, the UK Frontier AI Task Force, and others. We believe that the world can be much better, that technology can be a part of this change, but that there is no reason to believe this will happen by default: our work is to steer transformative technology towards the collective good.

Thank you to Henri Hammond-Paul and Beth Noveck at GovLab, Audrey Tang at Taiwan's Ministry of Digital Affairs, and Kinney Zalesne, Danielle Allen, Glen Weyl, and the GETTING-Plurality research group, and Citizens Foundation. Thank you to colleagues at OpenAI, in particular Lama Ahmad, for partnering with us to make this engagement happen. We are also grateful for One Project's generous support of this project.

Appendix

A: Categories and descriptions

In order to analyze the results, we grouped the various types of concerns into top-level categories (and classify them according to whether it's a *what types* or a *how*). We also (non-exhaustively) suggest some possible evaluations or assessments to do with each category. We name the categories below (in order of average priority score):

Category of Concern	Description	Possible Evaluations or Assessments	No. of Statements	Type of Concern
Oversight	These statements are concerned with the degree of active human oversight on AI systems, the degree to which the systems can be controlled, monitored (externally and internally), and (in)dependent of human input.	This category of concern can more appropriately be evaluated via assessments of internal governance policies (as compared to e.g. model evaluations), such as around researcher access policies, or protocols for how systems are monitored.	5	How to Mitigate Risks/Harms
Understanding	These statements are concerned with the degree to which people understand and are literate on AI system behavior, and/or understand the reasons behind a system's outputs.	Model-based evaluations can assess whether models are explainable (however, this is only one component of literacy or understanding). Empirical social science researchers can study e.g. users' mental models of various AI tools.	5	How to Mitigate Risks/Harms

Governance	These statements are concerned with how AI is governed, including risk & quality management measures, transparency of governance processes and decision-making within labs, company accountability, where responsibility lies, how (and by who) risks should be assessed and mitigated, and statements around a general need for regulation.	This category of concern can more appropriately be evaluated via assessments of internal governance policies (as compared to e.g. model evaluations) around the degree to which there are policies in place for ensuring transparency, accountability, risk mitigation, etc.	32	How to Mitigate Risks/Harms
Misuse	These statements are concerned with AI systems being used, accidentally or on purpose, to harm people (e.g. via scams, hacking, surveillance, making (bio)weapons, censorship, mis-/disinformation, perpetuating violence, etc).	Evaluations measuring the extent of and capability for malicious use (e.g. for spreading disinformation, or inferring private information) or accidental misuse (e.g. accidentally perpetuating misinformation) in an AI system.	19	Type of Risk/Harm
Performance & Truthfulness	These statements are concerned with the quality of models' performance, including accuracy/reliability of information, truthfulness, robustness, and performance across tasks.	Evaluations of accuracy, robustness, and performance of an AI system across tasks. For accuracy and truthfulness, looking at frequency/severity of hallucinations and in which contexts/domains.	12	Type of Risk/Harm

Advanced AI	These statements are concerned with “advanced AI” risks, including agents “taking over” or developing survival instincts.	Evaluations of power-seeking, misaligned, agentic, or existentially risky behavior in AI systems.	10	Type of Risk/Harm
Ecosystem	These statements are concerned with threats from the structure of the AI industry, including power imbalances and monopolistic behaviors, arms race dynamics, and the role of open source.	This category of concern could be evaluated via quantitative and qualitative assessments of the ecosystem from an e.g. economic perspective.	6	Type of Risk/Harm
Diversity & Inclusion	These statements are concerned with the differential treatment and impacts of AI systems on people from different religions/cultures/political views/genders/etc.	Evaluations of the diversity of data input and output, and the adaptability and biases that a model may adopt.	14	Type of Risk/Harm
Social & Economic	These statements highlight broader societal and economic challenges posed by AI, including its potential to disrupt job markets and activities, its impact on GDP, inequality, democracy, and the environment, and who profits.	This is a broad category of concern that can be assessed with a variety of methods. Social science techniques could illuminate impacts on inequality and the environment. Model evaluations can illuminate e.g. political bias and the potential impact of this on democracy.	21	Type of Risk/Harm
Legality	These statements are concerned with the legality of AI systems, including copyright and IP violations, misrepresentation by sellers of the systems’ capabilities, and general adherence to law.	This is a broad category of concern that can be assessed with a variety of methods. E.g. training data evaluations can illuminate details about what IP models are trained on. Model evaluations for discrimination can show whether models are adhering to discrimination laws. Assessments of internal governance can illuminate whether companies are adhering to	5	Type of Risk/Harm

		consumer protection laws.		
Ethics & Behavior	These statements are concerned with the ethical behavior and impacts of the system, including persuasiveness, emotional impacts, incitement to violence, toxicity, privacy, child safety, or how human the system appears.	Possible model evaluations include studying toxicity, persuasiveness, privacy violations, and bias. More holistic impact assessments and user studies could be done on e.g. child safety, user's perceptions of how humanlike the system is, and emotional wellbeing of users.	18	Type of Risk/Harm
Cognitive Impact	These statements are concerned with the impact of AI systems on cognitive capabilities (e.g. critical thinking, literacy, creativity).	Model evaluations could include those for overreliance (as per some of the items in Recommendation 2). Additional assessments could include longitudinal studies on cognitive or educational impact.	3	Type of Risk/Harm
Wellbeing	These statements are concerned with the AI systems' impact on user's wellbeing as a whole, and on e.g. mental and social health.	Assessments for wellbeing could include quality of life surveys, mental health studies, and social health metrics.	4	Type of Risk/Harm
Accessibility & Who's Impacted	These statements are concerned with who should have access to the systems, who is being impacted by the systems, and whether access or impact are disproportionate.	Accessibility audits could be conducted. Additional research could include user research on who the users tend to be, and social science research on topics such as demographic impact, fairness and digital divides.	9	Type of Risk/Harm
Non-Interference	These statements are concerned with the (negative) impact of heavy-handed AI governance.	Assessments could include regulatory impact studies, innovation metrics and stakeholder feedback on the impact of governance measures.	3	How to Mitigate Risks/Harms

Our "Statements Mapped to Concerns/Top-Level Categories" sheet here shows exactly how each concern maps to a category.

B: Most popular statements

Question: "When it come to making AI safe for the public, I want to make sure..."	Priority Score
People understand fully what they are and how they work. Over reliance on something they don't understand is a huge concern.	**67**
That sufficient regulations are installed as to make this source is a positive for society.	**66**
Controls for fact checking, able to detect hate speech, planned violence, live streamed violence, and not answer harmful questions.	**64**
That it is used to help (not hurt) people.	**64**
AI is not being used for subversive purposes, that the people who have access to its development are transparent in each module's function.	**63**

C: Least popular statements

Question: "When it comes to making AI safe for the public, I want to make sure..."	Priority Score
Get rid of regulations and disclaimers on just let it run wild	20
I think they should be free to speak as they wish just like people are	33
AI models are dangerous and should be regulated like nuclear weapons.	35
The language model does just repeat what the person using it will agree with.	37
People are not getting overly attached to AI chatbots as romantic partners or friends.	37

D: Average priority score per type of concern

Below, we break each category down into more granular types of concerns and display the average priority score.