Carnegie Mellon University Responsible AI

University of Pittsburgh Institute for Cyber Law, Policy, and Security

April 26, 2024

VIA ELECTRONIC SUBMISSION

Re: **RFI for Responsible Procurement of Artificial Intelligence in Government** OMB-2024-0004

This comment is submitted on behalf of Carnegie Mellon University Block Center for Technology and Society's Responsible AI Initiative and the University of Pittsburgh Institute for Cyber Law, Policy and Security in response to OMB's Request for Information concerning "Responsible Procurement of Artificial Intelligence in Government."

Over the past six months, our team of interdisciplinary faculty and researchers[1] at Carnegie Mellon University (CMU) and the University of Pittsburgh, representing expertise in computer science, policy, and law, have been leading an effort to understand the needs and challenges faced by U.S. cities in procuring AI products. To conduct this research, to date we've interviewed 17 municipal employees in six different U.S. cities, ranging in size (from populations <50,000 to 1 million+), operational budget, and region (including cities from the West Coast, South, and Midwest). Our goals are to understand cities' existing policies and procedures surrounding AI procurement, and to characterize the state of affairs surrounding responsible procurement of AI technologies. From these learnings, we plan to develop research and policy recommendations for the broader AI community and municipalities.[2] This research builds upon prior work from both institutions, including the Pittsburgh Task Force on Public Algorithms.[3] Our comment builds upon interview findings that will be discussed in a forthcoming public report. While focused on the needs of municipalities, many of our findings are applicable to federal procurement of AI products and services, including specifics of structuring responsible AI procurements, understanding public employees' relationships and negotiations with third-party vendors, and improving the public sector AI procurement marketplace writ large.

We commend the initiative from OMB in identifying key considerations and processes as they pertain to public sector AI procurement. In conducting interviews with city officials, we have encountered different levels of comfort and confidence in integrating AI into public sector services. The City of San Jose is a front-runner in this regard and has been the driving force behind the Government AI ("GovAI") Coalition, a group of over 400 public servants established to create shared guiding principles, policy resources, and vendor expectations. The GovAI coalition has done commendable work in generating deliverables,[4] including draft impact assessments and contracting language, ready for adoption and use by local, county, and state governments. We encourage OMB to review these products to inform federal considerations.

---

[1] Nari Johnson, Machine Learning PhD student at Carnegie Mellon University; Jessica Kuntz, Policy Director, University of Pittsburgh Institute for Cyber Law, Policy, and Security; Ravit Dotan, founder and CEO, TechBetter; Motahhare Eslami, Assistant Professor, Human-Computer Interaction Institute, Carnegie Mellon University; Hoda Heidari, K&L Gates Career Development Assistant Professor in Ethics and Computational Technologies, co-lead of Responsible AI Initiative and K&L Gates Initiative for Ethics and Computational Technologies at Carnegie Mellon University; Harrison Leon, master's student and Lingren Fellow, Carnegie Mellon University; Beth Schwanke, Executive Director, University of Pittsburgh Institute for Cyber Law, Policy, and Security; Elise Silva, Postdoctoral Associate, University of Pittsburgh Institute for Cyber Law, Policy, and Security.

[2] Our research methodology was approved by the CMU Institutional Review Board (IRB).

[3] Reporting of the Pittsburgh Task Force on Public Algorithms." University of Pittsburgh Institute for Cyber Law, Policy and Security. https://www.cyber.pitt.edu/sites/default/files/pittsburgh_task_force_on_public_algorithms_report.pdf

[4] Government AI Coalition, Deliverables. [link]

We further recommend that OMB and federal government agencies develop and deploy tools and procedures that can be readily adapted by subnational procurement officials. Our past research found that interviewed local and state employees desire federal guidance contextualized to the specific context of public sector procurement.[5] The federal government's purchasing power can also help incentivize a broader marketplace shift for vendors to commit to responsible AI practices and increased transparency. Additionally, strong federal procurement guidelines can serve to empower subnational governments in demanding more vendor transparency, remedying existing imbalances of power.

Below, we address questions 5, 6 & 10 raised in the Request for Information. Some responses include direct quotes from interviewed municipal employees (emphasis added). To preserve interviewee anonymity, we do not reference individual employees or cities by name.

Question 5. What access to documentation, data, code, models, software, and other technical components might vendors provide to agencies to demonstrate compliance with the requirements established in the AI M-memo? What contract language would best effectuate this access, and is this best envisioned as a standard clause, or requirements-specific elements in a statement of work?

One important dimension of compliance is determining which specific procurements involve AI. Our research suggests this task is more complicated than it may seem. Identifying "covered" AI applications is a prerequisite step that each agency must take to add them to the AI Use Case inventory and conduct necessary review processes (detailed in the AI M-memo Section 5). Many US cities have begun to institute similar review processes, in which municipal employees tasked with AI oversight must "flag" purchases with an AI component. <u>However, the majority of municipal employees we spoke to find it difficult to identify when a purchase involves AI.</u> In the absence of contractual language that requires vendors to *proactively notify* cities when they are using AI, many cities only learn that AI has been incorporated into an existing procurement when it is flagged by a *user* of the technology.

We first flag two common types of procurements where municipal employees struggled to identify when AI is being used: (1) new AI features integrated into *existing enterprise software*, and (2) AI as a key component of procurements *that did not involve explicit technology acquisition.* We subsequently offer thoughts on concrete steps that agencies and oversight bodies (i.e., the OMB) can take to retain transparency into all AI uses.

1. New AI features in existing enterprise software. Many cities shared that they often acquired AI effectively by accident, when new AI features were integrated or rolled into *existing enterprise software* that they had *previously purchased*. A common example is that some cities with existing contracts for word processing and related office suites noticed that there were new generative AI productivity tools being integrated into their software. When these new AI features were released, the city was often *not* notified by the vendor, and those tasked with overseeing AI governance were unaware until a user flagged the feature. One interviewee summarized the challenge of identifying these new AI features as follows:

"There's the gap of: we have all of this existing technology, thousands of applications and platforms and things that we have existing governing contracts with. So, as they do feature updates, or new pushes, or security patches, all this stuff just gets pushed. Now we have these types of AI that we probably maybe don't know about at this time, because we don't have standard AI contract terms saying, 'Hey, if you do something that meets this definition, you roll out a feature, you have to tell us'. We don't have those terms in place yet. So these things are coming into our existing tech stack. They're not going through traditional procurement, because the tech is already in use. It's hard."[6]

[5] "Studying Up Public Sector AI: How Networks of Power Relations Shape Agency Decisions", Kawakami et al., 2024. [link]
[6] Some responses include direct quotes from interviewed municipal employees (emphasis added). To preserve interviewee anonymity, we do not reference individual employees or cities by name.

Beyond the challenge of identifying these new AI features, another concern interviewees expressed was balancing their desire to ask "hard questions" of vendors (e.g., to understand the potential risks of new AI features), while maintaining existing relationships with vendors that provided enterprise software critical to city operations (e.g., their document management system). As detailed in the AI M-Memo Section 5.c.v.F, maintaining agencies' ability to "conveniently opt-out" from AI functionality – without any compromises to the functionality of the original product – is critical in such scenarios. We encourage agencies to include a similar contractual requirement for any software procurement that might someday incorporate AI functionality. This can include an additional requirement ensuring retention of any federal government data.

2. AI as a key component of procurements outside of traditional IT/software acquisition. One common assumption is that AI is only a relevant consideration for purchases of *technology* (e.g., software or online services). However, our conversations with municipal employees revealed how AI can still serve as a *principal component* in a procured service, *even when no technology is acquired* by the purchasing organization. As such, we urge agencies to conduct full due diligence to understand if AI is to be used to execute the scope of work, *for any procurement*, not only procurements that explicitly involve technology.

One illustrative example is from a city that procured a professional service from a vendor intended to support decision-making on allocating bond construction funding. The vendor presented the city with a report card that assigned each street a quality rating, with the understanding that the lowest-scoring streets would be prioritized for funding. However, the report card came under scrutiny when a city council member noticed that it recommended to "*invest in infrastructure [only] in some of [the city's] wealthiest districts*". After some research, the city discovered that the vendor was using a proprietary AI model to assign the grades to each street. Because this specific procurement did not involve an explicit acquisition of software – it was "*just a professional service contract for street indexing*" – the city was unaware that the tool even utilized AI. A city employee reflected: "*... that is kind of an illustrative example of how procurement of AI is not always as straightforward as, 'Oh, it's an AI tool that we are procuring'. In this case, it was a professional service. It was a company that was leveraging AI in their practices, and that wasn't clear to us.*"

This example makes clear how vendors' use of AI tools to fulfill a scope-of-work can still significantly influence critical decision-making processes, even in procurements that do not explicitly involve the acquisition of technology. As such, agencies should take adequate steps to ensure they proactively identify these uses of AI.

3. Recommendations. To help agencies identify "covered AI" for which vendors are required to provide access to technical components, we recommend that:

R1. Agencies include a standard contract clause requiring vendors to notify the agencies when they incorporate *new AI features* into existing procurements. The clause should include specific notification instructions. Agency's Chief AI Officers should be notified and immediately review existing purchases when new AI functionality is incorporated.

R2. When appropriate, agencies include a standard contract clause requiring vendors to implement and provide "opt-out" of AI functionalities introduced after an initial purchase. This opt-out should be at an organizational level, and, depending on the contract, might also include opt-out ability at an individual user level. The clause should include that the vendor must guarantee continued service and functionality of the original product or service, even if the organization/users choose to opt out from newer versions integrating AI, as per the AI M-memo Section 5.c.v.F. For example, organizations should be able to disable a new AI functionality added to their email client without losing any of the original email client functionalities. The clause should also include requirements on where and how the vendor must post the notice to opt out.

R3. Agencies mandate the same level of diligence on behalf of the Chief AI Officer to identify when a new purchase involves "covered AI", even for contracts that do not explicitly involve the acquisition of an AI system. To reduce the reviewing burden on behalf of the Chief AI Officer, each agency's procurement officials should be trained to identify and ask vendors if they plan to use AI to complete the scope of work. Agencies can support employees in this endeavor by providing clear definitions in lay language, and including specific examples of what AI tools can do (e.g., including popular generative AI tools like chatbots, but also examples of other systems such as predictive systems, robotics, vision or sensing technology, recommendation systems, decision support, etc.).

R4. For previously procured products and services, the above additions to standard contract language should be incorporated during award renewal when relevant.

The federal government has the opportunity to set an industry norm by instituting reporting requirements for vendors serving the federal government. Vendors can send the notices required by federal agencies to their other clients (e.g., municipal and state governments) at scale.

Questions 6. Which elements of testing, evaluation, and impact assessments are best conducted by the vendor, and which responsibilities should remain with the agencies?

Invariably, all AI quality control efforts demand collaboration between agencies and vendors. Below we discuss learnings from our interviews about how impact assessment and evaluation are organized and conducted by municipalities. We offer recommendations as to *when* in the purchasing process they should occur, and relatedly, the division of responsibilities between parties.

1. Regarding algorithmic impact assessments: Several US municipalities that recently began implementing AI impact assessments as a component of their procurement process[7] have found success conducting these assessments at the beginning of the design phase of a project, even *before* a specific vendor or AI system has been identified.[8] Similarly, we believe that federal agencies would benefit from conducting an early internal AI impact assessment *before* potential vendors are identified. Several academic works have argued in favor of conducting AI impact assessments as early as possible in an AI project's lifecycle,[9] as early assessments can help public sector employees identify and engage with impacted communities early pre-deployment. Early assessment can also mitigate costs by informing decisions not to pursue a project early on. In interviews, municipal employees mentioned that they used early impact assessments to inform specialized questions, solution requirements, and evaluation criteria in RFPs and subsequent vendor reviews. Early impact assessment can facilitate reflection on procedural mitigation steps that might be important requirements to ask of vendors during an RFP process. For example, an impact assessment might reveal the importance of retaining exclusive data ownership, which could then be added as a requirement in a solicitation.

We recognize that some questions in the impact assessment require vendor cooperation and, as such, can only occur after a specific AI system has been identified. Therefore the impact assessment should be re-visited *iteratively*, throughout the procurement process. To summarize, first, agency employees (such as the purchaser, intended users, and the agency's AI Governance Board) can document and review

[7] One example is the City of San Jose, who made several impact assessments public on their Algorithm Register.
[8] This is also the approach adopted by the Canadian Government in their 2023 Directive on Automated Decision-Making. They state: "The AIA should be completed at the beginning of the design phase of a project."
[9] See report from AI Now (2018), report from Data & Society (2021), and article from Kawakami et al. (2024).

responses to Parts 1 and 2 in the AI M-memo (e.g., reason through impacted stakeholders, failure modes, and risks to underserved communities), and use these responses to inform their solicitation and vendor evaluation process. The agency should directly involve (non-technical) stakeholders who hold expertise in thinking about how the agencies' work might impact vulnerable, under-represented, or historically marginalized communities in completing the assessment, as we elaborate on in our response to Q10. Once a vendor is identified, the agency and vendor can work collaboratively to address remaining questions in Section 3 in the AI M-memo (e.g., assessing "the quality and appropriateness of the relevant data", particularly for their model's training data). The vendor and agency should also revisit Parts 1 and 2 given the vendor's knowledge of the specific AI system(s) being considered (e.g., their knowledge of where their specific model might underperform).

The federal CIO Council has released an alpha version of an algorithmic impact assessment (AIA) instrument; however, the scope and objective of the CIO's instrument is unclear, and does not necessarily emphasize rights- and safety-related impacts, as prioritized by the AI M-memo. As such, we recommend that OMB publishes a base AIA instrument for safety- and rights-impacting AI, and that individual agencies maintain and publish their own instruments that are tailored to their sector. Maintaining and updating public-sector-bespoke instruments can foster innovation and empirical research on impact assessment, and help organizations experimenting with mandatory impact assessments share learnings and best practices.

2. Regarding evaluations: Vendors might proactively provide potential customers with their own bespoke evaluations. Below, we provide specific guidance on how agencies can better interpret the measures reported by vendors. We discuss two specific concerns: limitations on how evaluations on vendor-curated datasets will generalize to the agency's deployment context (2.1); and the validity of existing popular benchmarks (e.g., HELM) for generative AI systems (2.2).

2.1 Vendors' own evaluation datasets may not represent the deployment context: Often, vendor-curated datasets used for evaluation purposes, on which the vendor reports aggregate measures of performance, such as accuracy, will *not* reflect the population for which the agency intends to use the model. This concern is best captured in an example: one of the municipalities we spoke with had a sizable Vietnamese population and wanted to procure a machine learning (ML) tool to support translation services. The vendor assured the municipality that the system had been trained and tested with Vietnamese language. However, once they procured the tool, it quickly became apparent that the tool had not been trained in the particular dialect of Vietnamese used by residents and, as such, performed poorly. This example serves to illustrate that agencies should take caution in interpreting such aggregate metrics as a measure of how the model will perform in their *own deployment context*, which may differ from the dataset that the vendor used for evaluation.

Some concrete questions that agencies can ask to better understand vendors' evaluations are: 1) What data was used to conduct the evaluation? 2) From what population was this data collected, and when? 3) Does this population differ meaningfully from the deployment population: for example, are there groups of people in the *deployment* population that might be under-represented in the *vendor's* evaluation dataset?

2.2 Existing benchmarks for generative AI systems may not capture real world requirements: Beyond vendor-curated evaluation sets that are often used to evaluate *supervised* AI systems, a growing number of actors across academia and industry have curated their own public "benchmark" datasets that others (e.g., vendors) can use to evaluate the capabilities of generative AI. Examples include MMLU, NarrativeQA, LegalQA, and others[10]. Standardized benchmarks have several benefits, such as facilitating easy comparison across vendors.

However, one critical point for agencies to consider is that the tasks that these benchmarks evaluate are often detached from real-world uses. For example, while LLMs may be able to accurately answer multiple-choice questions from a bar exam (a common task in benchmarks),[11] high performance on such exams does not imply AI is fit for use by a public agency. Scholars conducting independent audits of AI systems procured by municipal governments have found significant flaws in their reliability and functionality: a procured chatbot service in New York City advised business owners to break laws and steal workers' wages.[12] Chatbots that achieve state-of-the-art performance on benchmarks, such as ChatGPT, still provide inaccurate and misleading information about the upcoming U.S. election.[13]

As such, we encourage agencies to ask the following questions to understand vendors' use of publicly available benchmarks for generative AI systems: 1) What specific capabilities (e.g., what types of knowledge – basic "reasoning"? knowledge of the US legal system?) does this benchmark assess? 2) Is there a gap between the capabilities targeted by the benchmark, and the agency's intended use(s) for the model? 3) How are the AI's generated responses scored? Against what reference (if any)? 4) How are summary statistics, such as aggregate measures of performance, calculated?

Evaluating and monitoring procured AI models to understand their performance *within* realistic deployment contexts requires collaboration between the procuring agency and the vendor. Moving forward, we discuss two concrete ways that agencies can work closely with and support vendors in their evaluation efforts. The first involves exploring short-term purchasing agreements with close oversight (i.e., "creating sandboxes") so that vendors can prototype their systems (2.3). The second focuses on creating use-case-grounded benchmarks for the public sector (2.4).

2.3 Prototyping & Sandboxes: Try before you buy! Several municipal governments have started experimenting with innovative purchasing pathways where vendors work closely with cities to prototype emerging technologies. One city that we spoke to implemented a new program where they pay vendors for short-term, small-scale "prototypes" that take place on the order of weeks to months. These prototypes equip city employees to better understand the potential benefits, limitations, and risks of emerging technologies. The prototype takes place under specific rules and conditions: for example, the prototype cannot access city data systems or infrastructure and the vendor is required to destroy any artifacts or data obtained from the prototype after the end date. The city referred to the set of rules and infrastructure they had created to enable these deployments as a "sandbox environment" for them to work with vendors to test and evaluate real-world use.

One interviewee described a short-term purchase of a resident-facing chatbot as an example of a prototype. The city found in a short-term pilot that the chatbot "*started hallucinating and interacting with people in really unpredicted ways*", which "*sparked an interesting conversation with the community about the sorts of risks of AI tools in the public space*". When reflecting on how this pilot revealed the potential flaws of purchasing such a chatbot, the interviewee stated: "*This is exactly why we need a [prototyping] program like ours, so that we create a safe space to test these things out and explore their capabilities and understand what it would actually mean in practice. You learn so much more by just deploying this, than you do by trying to plan out every detail and make it perfect in advance.*"

[10] The LLM benchmarking space is rapidly evolving in response to criticism that existing benchmarks are often flawed and contain scoring errors. Our critique applies broadly to the majority of question-answering based benchmarks in the field. See the HuggingFace OpenLLM leaderboard for an example list of metrics.
[11] See the list of HELM scenarios.
[12] "NY's AI Chatbot for small businesses suggests them to break laws, steal wages". Firstpost, 2024. [link]
[13] "Opinion: AI doesn't have all the answers – especially this election season". LA Times, 2024. [link]

As the interviewee stated, short-term prototype purchases can inform a future procurement process for the technology (e.g., RFP requirements or contractual terms): "*We're generating stronger procurement pathways, so that by the time we get to that contract point, we've addressed all these issues through testing, instead of the alternative which is issuing an RFP for a cool technology we don't understand.*"

We encourage federal agencies to consider utilizing sandbox environments or other forms of short-term technology evaluation agreements, to allow observation and use of an AI system within its intended deployment context, *before* proceeding with a long-term purchase. A period of government user testing in the real-world deployment context can identify needed mitigations prior to proceeding to purchase, or even prevent the purchase of a system that might cause harm or have efficacy concerns in its deployment.

2.4 Public benchmarks for the public sector. We recommend that federal agencies invest in the creation of use-case-grounded, public-sector-specific benchmarks. Specifically, given a specific use case for an AI model (e.g., a chatbot deployed to answer residents' questions about government services),[14] federal agencies can create a public benchmark to compare different vendors' performance across a set of tasks. Beyond curating an evaluation dataset and scoring system, creating such benchmarks would also entail developing infrastructure for vendors to submit candidate models for evaluation. One concrete example of a use-case-specific benchmark is the NIST Face Recognition Technology Evaluation (FRTE)[15]. Vendors wrap their (potentially proprietary) models behind NIST's public API, run software that NIST has provided to calculate the model's predictions on the evaluation dataset, and upload their results for their models to be listed on NIST's public leaderboard.

Standardized, sector-specific evaluations have the distinct benefit of allowing purchasers to directly compare competing models – comparisons that are at present difficult because different vendors often use incomparable evaluation processes (i.e., different evaluation datasets). City employees we interviewed expressed that they would like to see standardized public benchmarks when purchasing AI-driven license plate readers, voice recognition, and translation algorithms.

While benchmarking has several benefits, other types of evaluations might also be necessary to anticipate potential negative impacts, especially in high-stakes scenarios. When appropriate, agencies should consider additional evaluations beyond benchmarking, such as red-teaming, adversarial evaluations, and additional testing to discover the model's failure modes.[16]

Question 10. How might OMB ensure that agencies procure AI systems or services in a way that advances equitable outcomes and mitigates risks to privacy, civil rights, and civil liberties?

Broadly, the risk mitigation processes and requirements for federal agencies in *external* procurement of AI systems/services should align with those for *internal* development of AI systems/services, as specified in the AI M-memo. We discuss three additional recommendations below.

10.1 Participation by Multiple Internal and External Stakeholders for High-Risk Systems: One significant method assisting municipalities in anticipating and understanding the risks of procured technologies to impacted groups, is to involve internal and external stakeholders in critical early planning conversations, such as the impact assessment process. We urge federal processes to consider doing the same, particularly for high-risk systems impacting privacy, civil rights, and civil liberties. Stakeholder engagement efforts may include internal staff, public health officials, civil right advocates, or others tasked with thinking about how the agencies' work might impact vulnerable, under-represented, or historically marginalized communities.

[14] https://www.sanjoseca.gov/home/showpublisheddocument/109818/638460264984830000
[15] https://pages.nist.gov/frvt/html/frvt11.html
[16] See [Feffer et al. 2024] for a review of red-teaming approaches, and [Cabrera et al. 2023] for a survey of methods to surface previously unknown model failures.

One municipal Chief Data Officer spoke about the importance of involving their city's Chief Diversity Officer in the impact assessment process for technologies such as facial recognition and other policing technologies that impact residents' civil rights and liberties: " *[...] I know that bias exists in the data, but that's not my area of specialization. And on the DEI side, they're not experts in technology [...] but their perspective can inform how we're assessing that. I don't think that there is any way to replace or have a stand-in for including their perspectives in the conversation. I don't know how cities are going to look at bias and equity, if you don't have people who specialize in bias and equity in the room with you. I think they need to be allowed to ask questions.*"

In addition to internal government stakeholders, integration of *external* stakeholders, particularly civil society groups and representatives from impacted populations, can provide valuable understandings of risks and potential mitigations.[17] One municipality successfully leveraged existing constituent communication channels (e.g., by attending neighborhood association meetings) to elicit feedback from impacted communities. Federal agencies can similarly look to leverage existing channels to engage residents in policymaking.

Additionally, agencies should bear in mind that successful mitigations of privacy, civil rights, and civil liberty risks are not guaranteed by initial steps taken during a procurement process. Monitoring and stakeholder consultations should be ongoing and iterative.

10.2 Ensure LPTA Does Not Undermine Procurements of Responsible Systems: We urge OMB to limit federal agency use of the lowest price technically acceptable (LPTA) process in AI procurement. Federal Acquisition Regulation 15.101-2 lays out the terms under which LPTA should be used and states that LPTA "is appropriate when best value is expected to result from selection of the technically acceptable proposal with the lowest evaluated price." We would argue that ensuring the selected service provides sufficient transparency to proactively screen for bias and other harms may be a higher priority than price. The rule of thumb should remain: *if purchasing unbiased, privacy-preserving, and compliant AI is cost-prohibitive, it is better not to utilize AI in that instance.*

It bears further reflection as to whether fairness and privacy should be considered as a floor (a minimum set of standards that all tools must meet for consideration), or whether agencies should give preference to tools that more effectively address potential risks concerning fairness and privacy. The latter approach could help to foster a virtuous circle whereby vendors outcompete one another, fostering innovation in mitigation strategies and methods, disclosure, audit processes, red teaming, etc.

10.3 Expand Applicability Beyond Formal Procurements: Finally, OMB should also thoroughly consider the range of ways in which acquisition can take place outside of the formal procurement process. In the case of municipal government, this was found to occur via purchasing cards, research collaborations, donations, piggyback contracts, and other contracts under a set dollar amount exempt from full procurement procedures. Such processes are intended to foster efficiency, but risk creating a backdoor whereby AI tools enter into use without comprehensive evaluation. Risk can still be introduced in acquisitions where money does not change hands. *OMB should therefore consider how best to apply standards and regulations developed for AI procurements to other forms of acquisition.*

[17] Creating opportunities for meaningful participation may also require investing in creating new social infrastructure. See [Young et al. 2024] Sections III, IV, and V for key considerations for and example case studies of participatory AI.