Public Procurement for Responsible AI? Understanding U.S. Cities' Practices and Needs

Nari Johnson[1] **Elise Silva**[2] **Harrison Leon**[1]
Motahhare Eslami[1] **Beth Schwanke**[2] **Ravit Dotan**[3] **Hoda Heidari**[1]
[1]Carnegie Mellon University
[2]University of Pittsburgh Institute for Cyber Law, Policy, and Security [3]TechBetter
{narij, meslami, hheidari}@andrew.cmu.edu
{beth.schwanke, elise.silva}@pitt.edu
ravit@techbetter.ai

Abstract

Most AI tools adopted by governments are not developed internally, but instead are acquired from third-party vendors in a process called *public procurement*. While scholars and regulatory proposals have recently turned towards procurement as a site of intervention to encourage responsible AI governance practices, little is known about the practices and needs of city employees in charge of AI procurement. In this paper, we present findings from semi-structured interviews with 18 city employees across 7 US cities. We find that AI acquired by cities often does not go through a conventional public procurement process, posing challenges to oversight and governance. We identify key types of challenges to leveraging procurement for responsible AI that city employees face when interacting with colleagues, AI vendors, and members of the public. We conclude by discussing implications for AI and policy researchers.

1 Introduction

Artificial intelligence is increasingly utilized in the public sector to automate bureaucratic process and workflows, and assist decision-making processes that impact residents [77, 56, 25, 35, 97, 52, 57]. Often, public-sector AI systems are not developed in-house, but are purchased from external third-party vendors through a process called "public procurement" [91, 72, 58]. In fact, in a 2023 opening statement for the full committee hearing on AI and procurement, U.S. Senator Gary Peters stated that "over half of the AI tools used by federal agencies have been purchased from commercial vendors" [70]. Experts estimate that this number is even higher at lower levels of government, such as state and local governments that are even less likely to have internal expertise to develop AI [83, 84, 64]. Thus, most public-sector AI systems used today are developed by and acquired from *private vendors*.

A growing body of academic and advocacy efforts have pointed out how AI systems procured in the public sector have predominantly targeted narrowly defined notions of efficiency and performance enhancements, resulting in adverse effects that disparately impact marginalized communities [35, 82, 92, 47, 51, 18]. Facial recognition AI technologies, for example, have faced sustained criticism due to concerns about civil liberties, racial biases and privacy violations [98, 48, 2]. Despite these concerns, in 2022, a U.S. government watchdog revealed over 20 federal law enforcement contracts with private sector companies that either specialized in facial recognition or had awards related to its use, with a total budget exceeding $7 million [80].

While such incidents have exposed flaws in individual AI systems, they highlight deeper issues in how AI is acquired, used, and governed in the public sector. The AI procurement process encompasses decisions of which AI tools to ask for, adopt or reject, and the manner in which they are developed and

2nd Workshop on Regulatable ML at NeurIPS 2024.

deployed: decisions of critical importance for communities who may be harmed by AI. Such decisions not only influence the performance and risks posed by AI systems, but also play a significant role in shaping broader governance practices and ethical standards by which AI operates in the public sector. Mulligan and Bamberger [64] argue that studying this procurement process is of dire importance as governments increasingly adopt algorithms that have irreversible and life-changing impacts on residents' lives. When procured algorithms are used to make critical bureaucratic decisions such as who to jail or who to separate, scholars argue that procurement *is policy*: a commitment to replacing human discretion with the policies and values that algorithms embed [38, 13, 43, 64].

In response to increasing incidents of harm caused by public AI [62], experts have called for governments to adapt their existing procurement processes to be enable the assessment and governance of risks and complexities unique to AI [83]. For example, in the United States, the federal government has released and started to implement the "AI M-Memo" [6, 89], a landmark policy that outlines mandatory steps all federal agencies must take to promote the "responsible procurement of AI technologies". Similarly, a few months ago local governments announced the formation of the Government AI ("GovAI") Coalition [9, 34, 81], a group composed of over 1,000 members representing 350 participating U.S. governments founded to "give local governments a voice in shaping the future of AI".

Thus, in a time when AI procurement has become a pressing matter of policy attention, we believe that empirical research—to understand the challenges government employees face when attempting to incorporate responsible AI considerations into their procurement practices—can help inform policy development and implementation. To date, there is a dearth of empirical research focused on understanding governments' AI procurement practices. To address this gap, this work builds on the burgeoning efforts across the United States to assist governments in procuring AI and investigates how city employees are approaching the procurement of AI systems. In this workshop paper, we present selected findings from semi-structured interviews with 18 city employees across 7 U.S. cities, aiming to address two key research questions:

- **RQ1. Characterizing existing AI procurement practices in public sector:** What practices do city employees follow to acquire AI products and services?
- **RQ2. Understanding challenges and desires to procure AI responsibly:** What challenges do city employees face through the AI procurement process, and what are their needs to address those gaps and challenges? What concrete resources can support them to overcome these obstacles?

We find that city employees face several key challenges in implementing responsible AI best practices, in part due to complexities in their relationships with AI vendors. Cities often did not participate in the development of procured AI systems, and instead were often sold AI systems developed to be used off-the-shelf. City employees also often encountered vendors' trade-secrecy claims, which posed barriers to adequately assessing risks posed by procured AI. We conclude by discussing implications for AI and policy researchers.

2 Background

While emerging AI technologies are new, governments must acquire *all* goods and services - including AI - using established *public procurement processes* that lay out required steps before a purchase can be made. We begin by introducing key components of this procurement process, and describe how they may be used to procure AI. We also present an extended discussion of past scholarship on public-sector AI and AI procurement in Appendix A.

A Public Procurement Primer for AI The term "public procurement" generally refers to the processes governments use to bring in goods and services that are developed externally [91, 72, 58], often involving paid transactions with third-party organizations.[1] In this work, our core focus is on studying public procurement practices in the United States, particularly for *local* (city) governments.[2]

[1]We note that as described by past work [72], there is no single precise agreed-upon definition for what it meant by the term "public procurement" – rather, the definition is "muddled" and varies across contexts. See Appendix B for a more detailed discussion of definitions.

[2]All of the cities we interviewed qualify as cities, but the procurement processes we outline here are also applicable for other types of US local government, *e.g.,* counties, municipalities, etc.

While procurement laws, organizational structures, and activities vary across different cities [58], they all specify common steps that take place in a *conventional procurement process*. These steps were designed to be applicable for *any* good or service, including pencils, school buses, and technologies, including those that have AI [28, 45, 41]. We provide a brief sketch and introduce key terminology necessary to understand this procurement process, and direct the readers to [84] for a more comprehensive review.

1. **Planning.** The procurement process begins with *planning* when public-sector employees identify a potential need or application for procured goods or services. For AI, the planning phase might involve identifying a context or use-case where employees believe that AI might be appropriate [67, 53].
2. **Solicitation.** Once an employee has decided that they are interested in using an externally-developed solution for their need, they then begin a *solicitation*, a competitive process to select a vendor. One type of solicitation is a *Request For Proposal (RFP)*, a structured process where a government outlines their needs, expectations, and desired outcomes. Interested vendors then submit detailed proposals that comprehensively address these requirements [5].[3] For AI, solicitations may include specific requirements and criteria desired of the procured AI solution [84]. AI vendors may also be invited to give a live demo of their product [1, 71].
3. **Review & Award.** In the *review and award* phase, cities evaluate vendor proposals using score sheets and other established processes. For AI, this phase might involves assessing how well the tools adhere to responsible AI standards, and deciding the level of risk that will be tolerated [79].
4. **Contracting.** Once a vendor is selected, the city and vendor create a *contract* that specifies legally enforceable obligations for both parties, such as the agreed price, statement of work, the vendor's support responsibilities, and an outline of how disputes will be resolved. This phase typically involves a *negotiation* involving activities such as *red-lining* (negotiating contract clauses). For AI, relevant contract terms may spell out expected AI risk management practices, such as regular performance monitoring, or procedures on how to respond to incidents where AI causes harm [32, 9, 40].
5. **Deployment & Use**: The deployment phase is when the procured AI solution is adopted and used by the city. For AI, this phase may involve training frontline workers and users who will consume AI outputs [54], and continued oversight and monitoring of AI risks and performance [9].

In our study, we return to these five steps to examine city employees' experiences procuring AI within a conventional purchasing process, and whether they made special adaptations for AI. Importantly, while much past work typically describes these five steps, we do not assume that all AI acquisitions are procured using this process. In doing so, we hope to characterize any differences between the conventional procurement process, versus the steps city employees actually took in specific real-world AI procurements.

3 Methods

Over a period of 6 months (from December 2023 to June 2024), we leveraged semi-structured interviews to understand city employees' needs and challenges surrounding AI procurement. We interviewed 18 city employees (described in Table 1) across 7 U.S. cities that varied by region and size. Participating cities represented all four major regions (Northeast, West, Midwest, and South) defined by the U.S. Census Bureau [7]. Participating employees included both department leaders (*e.g.,* Chief Technology Officers) tasked with providing strategic guidance and making decisions on behalf of their department, and other employees whose day-to-day responsibilities involved managing technology procurement.

Participating cities for the sample were recruited using both convenience and snowball sampling: we began by contacting city employees in our professional networks who had demonstrated a past interest in AI governance. We also cold-emailed or were introduced to employees at other cities that were referenced in our conversations with others. We intentionally selected and invited cities

[3] Beyond RFPs, there are several other types of formal solicitation processes, such as a Request for Information (RFI), Request for Bid (RFB), Requests for Quotation (RFQ), and others. See [10] for a more comprehensive review.

	Department	City Size	Title
p1	Innovation	Medium	Chief Data Officer
p2	Information Technology	Large	Senior IT Manager
p3	Information Technology	Large	Privacy Program Manager
p4	Information Technology	Small	IT Business Relationship Manager
p5	Information Technology	Large	Privacy Specialist
p6	Information Technology	Small	Director of IT
p7	Information Technology	Large	Chief Privacy Officer
p8	Management & Budget	Medium	Sourcing Specialist
p9	Innovation	Medium	Innovation Specialist
p10	Information Technology	Large	Privacy & AI Analyst
p11	Information Technology	Medium	Chief Technology Officer
p12	Human Resources	Small	Talent & Culture Program Manager
p13	Information Technology	Medium	Chief Data & Analytics Officer
p14	Management & Budget	Large	Director of Procurement
p15	Information Technology	Large	Chief Technology Officer
p16	Innovation	Large	IT Policy Director
p17	Information Technology	Large	Vendor Manager
p18	Innovation	Large	Chief Information Officer

Table 1: An anonymous description of participating municipal employees. Titles were modified to preserve anonymity. Small cities have under 200,000 residents, medium cities have 200,000 - 500,000 residents, and large cities have over 500,000 residents.

that represented a wide range of maturity surrounding AI (*e.g.,* whether or not they had privacy or AI-focused personnel or had adopted any public-facing AI policies). We invited 8 total cities, and 7 agreed to participate in the study.

To recruit participants, we used snowball sampling to ask our initial contacts at the city to introduce us to other eligible employees. Participants were eligible if their present role was involved with technology procurement or governance in their city. As shown in Table 1, the majority of interviewed employees worked in technology-focused roles in their city's IT or Innovation departments. We also were introduced to and spoke with specialists in vendor relations and procurement to provide a broad perspective on procurement processes, and one human resources representative who had conducted organizational training on AI. To maintain confidentiality and protect participant identity, all data was anonymized and presented in an aggregated form, with sensitive quotes excluded or paraphrased where necessary.

Semi-Structured Interviews We adopted a semi-structured interview approach to allow flexibility in discussions, enabling participants to express their thoughts freely and spend more time discussing phases of the procurement process that are closest to their responsibilities and expertise, while covering essential topics predetermined by the two RQs. Interviews ranged from 60 to 90 minutes, and the interview protocol included three sections. First, we asked participants about their background, such as their current role and work responsibilities relating to AI. We defined "AI" to participants as "any machine-based system that can make predictions, recommendations, or decisions"[4], and provided examples of qualifying systems such as facial recognition, resume screening, and chatbot technology. Second, we asked participants to walk through how an example procurement involving AI would occur in their city, paying particular attention to any differences between a standard technology procurement. Our goal was not to impose structure on participants' descriptions of procurement, but rather allow them to describe how they personally view the procurement process. In the final section, we asked participants to reflect more deeply on their perceived challenges, needs, and desires to improve the AI procurement process. The study was approved by a university Institutional Review Board (IRB), and participants provided informed consent. We include our complete interview protocol in Appendix C.

To preserve anonymity of participating employees, we assured interviewees that their participation was voluntary, they could decline to answer interviewer questions, and their responses would be kept anonymous. For sensitive or potentially identifying interview quotes, we exclude participant IDs to

[4]This definition is adopted directly from the OECD [39].

preserve anonymity. When appropriate, we use the "x" character to omit exact dollar amounts to preserve confidentiality.

Qualitative Analysis We collected 23 hours of interview audio which were transcribed and coded by four team members, including the principal investigator. We adopted a bottom-up thematic analysis approach [21] to analyze interview transcripts. Each transcript was open-coded by two authors, who met regularly to discuss each transcript and resolve any differences in interpretation [61]. The process was iterative, including regular discussions to adjust coding strategies, group codes into higher-level themes concerning employees' practices and needs, and refine the coding schema. In total, we created 305 unique codes, which we grouped into 43 higher-level themes. We discuss methodological limitations of our study in Appendix E.

In what follows, we discuss selected findings that we believe are of key relevance to the NeurIPS Regulatable ML community of AI and policy researchers. For an extended discussion of our findings and implications for other actors, we refer the reader to our upcoming manuscript.

4 Selected Findings

Cities are increasingly adopting externally-developed AI. We found that all seven of the cities that we interviewed shared that they had *already started to use third-party AI technologies* developed for a wide set of intended users and goals, *e.g.,* to aid law enforcement, inform urban planning decisions, assist bureaucratic decision-making, facilitate resident communications, and increase workplace productivity. In Appendix D, we describe each of these use cases in further detail and provide example real-world AI solutions mentioned by participants. Notably, almost all of the cities we interviewed *did not have the capacity to develop their own AI solutions internally*, motivating their need to adopt AI developed outside of their organization, with the exception of two large cities that were experimenting with developing their own AI solutions with their own IT workforce.

Some cities, but not others, have adapted their purchasing practices for AI. We found that three out of seven cities that we spoke to had already introduced explicit changes to their existing procurement practices for AI. The remaining four cities had not changed their processes, and assessed AI similarly to "any other technology" (P6), applying broad criteria for purchasing software such as "threats to cybersecurity" (*e.g.,* data breaches), "usability", and "inter-operability with [existing] enterprise systems" (P1, P4, P8, P18).[5] In contrast, participants in the three cities that revised their processes found these broad criteria insufficient to address novel risks posed by AI technology. For example, one procurement specialist (P17) reflected on how in contrast to traditional software procurements, which were "very contained algorithms", data-driven AI posed novel risks due to risks of inaccuracy, privacy risks and "human biases encoded in [training] data", inscrutability, and changes in behavior when models are "refreshed" (re-trained). Another participant (P18) called attention to the environmental impact of training and using large language models, a consideration they did not typically consider for the average software procurement.

Interestingly, we found that the common step that interviewed cities took to improve their AI governance was developing new *usage policies* for how city employees should or should not interact with publicly available generative AI tools, such as ChatGPT. Common advice included not entering private city data into chatbots, using city (not personal) accounts to discuss city business, reviewing AI outputs before using them, and disclosing when AI was used to generate content. In many cities, these generative AI usage policies were the government's first (and sometimes only) action taken regarding AI. While all seven interviewed cities had already or intended to release generative AI usage policies, these policies were often narrowly-scoped. For example, such policies only covered *generative* AI tools, and were not applicable to other types of AI systems, such as predictive AI [96]. Similarly, such policies focused on individual employees' usage of AI tools, rather than the broader *acquisition* processes by which AI tools *came to be in use*.

Cities that had changed their processes for AI could cities often a separate "AI review" process overseen by experts in technology, that occurred outside of conventional procurement processes (*e.g.,* led by an IT rather than a procurement department). When an employee wanted to acquire an

[5] Department leaders in three out of these four cities shared that they *intended to* change their practices for procurements involving AI, but had not yet decided how to do so.

AI solution, they initiated an AI review by completing a form or "opening a ticket". IT employees trained to assess the risks of the AI then would use information from the employees' request to conduct further reviews. AI reviews often involved AI-specific risk assessments, vendor reporting requirements, and negotiations to include additional AI-specific contract terms. We describe each of these components in more detail in Appendix F. In the following paragraphs, we expand on challenges that these cities faced when implementing these interventions.

Many AI acquisitions occur outside of conventional procurement processes. In Section 2, we described a classic procurement process as it is described in the literature. Our interviews, however, indicate that AI procurement often doesn't take this classic route, often skipping centralized planning, solicitation, competition, and contract negotiation phases. Participants pointed out that due broader shifts in the AI landscape, namely the availability of low- and no-cost AI tools (P1, P12, P13, P16), many AI acquisitions did not involve a competitive solicitation (*e.g.,* no RFP) because they were under cost thresholds that would require them to do so.[6] Local procurement law specified that municipal employees could make purchases under a certain dollar amount at their own discretion, using a government-issued purchasing card (sometimes called a "p-card"). Types of AI tools that fell under cost thresholds included free online services (*e.g.,* chatbots), services with paid subscription models (*e.g.,* coding assistants), or AI donated through academic collaborations, foundations, or from for-profit companies.

In many cities, AI acquisitions that occurred under cost thresholds "didn't have to go through procurement", and thus fell outside the scope of existing accountability structures for government purchasing. For example, one department leader reflected on how acquisitions under cost thresholds were particularly difficult to govern or even be aware of until after they had been purchased:

> "[For purchases] below $x0,000, there's few oversight or regulatory mechanisms to control, or even have visibility of what departments do. We can go back through our financial data to say, 'Oh, this money was spent on this procurement', but it's not routed through a centralized control mechanism." (P16)

The participant was particularly "concerned" about employees' use of free generative AI technologies after an experience where they learned that an employee started to use a free transcription tool that did not have "a consensual model for data collection". The participant reflected on the hidden costs of free AI tools:

> "If you are not paying for it, you're the product. We have to be mindful about the extractive capabilities of these tools that can be free, but are at risk of us divulging resident information, possibly more secure information as well." (P16)

Cities with AI reviews required employees to complete an AI review *before* procuring or using an AI system, regardless of cost. However, in cities without AI reviews, there were often no applicable processes to oversee free or low-cost acquisitions. In effect, individual users could purchase or begin to use free and low-cost AI tools (*e.g.,* with prices under $x00 per month) at their own discretion, without notifying others in their city.

"Is that hallucination thing for real?": Participants desire focused training on AI risks. Participants located in cities that had yet to develop AI-specific components of their procurement process expressed anxiety and a lack of confidence in their ability to assess AI solutions. When discussing what they perceived to be significant risks posed by AI systems, several participants discussed concerns about data security, data ownership and retention, and data privacy. However, a much smaller number of participants mentioned risks posed by (in)accuracy [74], (un)fairness [20], adversarial robustness [37], lack of transparency or explainability [14], contestability and recourse [65], or broader societal impacts, *e.g.,* to labor or the environment [88, 59]. For example, one participant mentioned that their team did not consider risks posed by "hallucinations" in a recent procurement of an AI chatbot service, and that risks due to inaccuracies were "not part of the conversations" they had with the vendor.

[6] These cost thresholds are specified by both state and municipal procurement law that applies for all purchases beyond technology, and varied across participating cities.

"Information is typically black-boxed": Vendor secrecy & obfuscation. While participants tried to leverage their city's purchasing power to ask more from vendors, *e.g.,* to provide basic information about their AI system or amend their contracting terms, many participants found that vendors were unwilling to amend their position. With some exceptions, participants repeatedly felt that they lacked leverage in advocating on behalf of their city. AI vendors frequently refused to disclose information that reviewers requested, claiming that the information was *proprietary* (*i.e.,* protected as the vendor's intellectual property). One participant explained how vendors' refusal to grant cities access or provide basic information about their system limited the participant's ability to make informed purchasing decisions:

> "What we need to perform a risk assessment is intimately tied to the [data] models that power AI systems, which most vendors treat as proprietary. So, having access to the model, which is the engine of how the AI tool is working, knowing the sources of training data that are being used, having information on the accuracy of the AI [...] this information is typically black-boxed." (P2)

Participants described other tactics beyond invoking IP that vendors used to avoid answering their questions, such as "ghosting" (not responding to e-mails) (P5), "deflecting" (P10), or simply stating that they cannot answer (P5, P10, P13). Other denied requests for information included questions about the presence of copyrighted content in a model's training corpus (P3, P5), whether data collected from employees' interactions with the AI would be used to train the vendor's models (P1, P17), and disaggregated performance measures of the model's accuracy, *e.g.,* across different demographic groups (P2, P7, P10). In response, several public sector employees with technical expertise shared that they conducted their own independent evaluations of third-party AI systems themselves (described further in Appendix F).

"Few companies are willing to do boutique AI models": Lack of customization. Several participants valued the ability to customize the AI services that they procured to the unique context and needs of their locality. However, participants shared that they often were not consulted or involved with the design or development of procured AI solutions. Instead, the majority of procured AI systems were designed to be deployed *off-the-shelf*, without being customized to (*e.g.,* trained or fine-tuned using data from) each city. One participant used the term "turn-key" to describe this type of vendor business model:

> "[Vendors] are like, '*We just want to scale our business model and get out of the game*'. To do that, it has to be this turn-key thing. Very few companies are willing to do 'boutique' AI models where they're taking your specific dataset and training their model off of that. " (P2)

Participants also wished that vendors modify their AI models to apply technical mitigations to reduce potential societal harms. As one example, a participant whose background was in data privacy discussed a positive experience where the participant worked with a vendor who developed AI models to count the number of people using city facilities:

> "We were like: '*Whoa, why are we just watching people?*' So we worked with the vendor to ask: Do we actually need this video on, or can you blur it, or make it a heat map? What functionality can you give the city so we don't have to literally watch humans walk in, when all we need is a count number?" (P3)

In this procurement, the vendor applied the city's requested mitigations and trained an AI model to detect people entering a building using alternative data sources, instead of a live video feed. This example illustrates the importance of implementing technical mitigation steps (*e.g.,* changing the form of the data given as input to a predictive model) to manage and reduce risks.

"What is acceptable here?": Determining organizational AI risk profile & risk tolerance Due to their varying levels of AI preparedness, expertise, and capacity, cities wanted to pick and choose the specific steps they would take to conduct their own risk assessment. Participants across cities repeatedly expressed uncertainty and confusion about determining their organization's *risk tolerance*, *i.e.,* how much risk they were willing to accept. Some cities established hard ceilings on certain types of risks by instituting minimum "red line" requirements [76] of procured AI systems. But sometimes,

participants struggled to find AI that met their requirements. For example, one department leader recently instituted "language in [their] city policy that city officials had to make a reasonable effort to ensure [AI] use was not violating existing intellectual property laws". In response, a generative AI vendor told the city that this requirement disqualified them from consideration. While the requirement worked as intended to protect the city from potentially using illegal software, the participant wondered it would disqualify most eligible vendors:

> "We had a number of conversations about that – in particular, are our standards too high? Or is the technology simply too risky or problematic for us to use effectively?" (P15)

In contrast to red line requirements, participants also had to make ad-hoc judgment calls after collecting relevant information from vendors, such as performance metrics. Some participants struggled to determine what values of the metric were good enough, or "set the line" (P9). One participant discussed how their department had trouble interpreting the values of the metrics reported by vendors when making decisions about whether to move forward with a purchase:

> "We ask some sort of question: '*What's your R-squared value*'? And how do we know if [what is reported] is good? Someone needs to be able to say, is that good or not good? Like, have that kind of technical acumen to say what is acceptable here in terms of accuracy, error rates, thresholds, or whatever." (P3)

Other participants also expressed enthusiasm for clear guidance and thresholds when interpreting measures of AI risk, noting that such guidance might be most impactful for "small jurisdictions that just don't have the capacity [to conduct AI reviews]" (P7). The participant conceptualized this guidance as a consistent "stamp of quality" for AI that could institute a minimal set of requirements, *e.g.,* for performance and non-discrimination. Yet, some participants surfaced the complexity of instituting uniform risk tolerance standards across their organization given that individual employees might have differing preferences for what risks to accept. As put by one participant tasked with data privacy reviews, "my risk threshold probably looks very different than everyone else's threshold" (P3). Thus, organizations will need to strike a balance between establishing clear thresholds when appropriate, while acknowledging that in some scenarios, individual employees may have different perspectives and preferences for where to draw such red lines.

5 Discussion

Governments today struggle to anticipate and mitigate harms caused by procured AI technologies. However, there is lack of scholarship on what government employees need. In interviews across seven cities, we consistently found that cities struggled to uphold private vendors to responsible AI best practices. City employees were uncertain of how to assess information such as performance metrics provided by vendors, if such information was even provided at all. Procured AI often was not trained on nor evaluated on city data before used in-deployment. Our findings point to the uniqueness of understanding the procurement relationship for responsible AI. How can researchers and policymakers support city employees, who are often better-positioned to represent the interests of their constituents than private vendors, in being accountable over AI technology that they did not develop, and often face fundamental barriers in accessing? In what follows, we highlight a few broad directions and provocations for researchers and experts in AI and policy to consider:

- Most existing infrastructure and tools to promote responsible AI best practices assumes that the user is also the *developer* of the model (*e.g.,* and has full access [23]). However, our findings show that vendors often deny cities access to their model, claiming that it is protected as their IP. One particular research direction of interest is the creation of tools that would allow non-expert users (*e.g.,* city employees involved with making purchasing decisions), with varying levels of access to a third-party AI system, to understand and evaluate AI behaviors [93, 46, 95].
- Data-sharing between vendors and across cities often does not occur. As such, there is a lack of public datasets available for common public-sector AI use cases, such as 311 translation services [9]. As a result, individual vendors often report the values of performance metrics calculated on their own proprietary datasets, making comparison across vendors difficult.

Researchers can support cities by creating measurement methodologies and evaluation infrastructure for public-sector AI, *e.g.,* both standardized benchmarks and field testing guidance. For example, one participant shared that they collaborated with a university to translate policy objectives into quantifiable "measures of success" for a generative AI project. Academic researchers can similarly design empirically-informed resources and tooling to support evaluation and measurement. Researchers can aid cities in creating standardized benchmarks to enable consistent comparison of performance across vendor offerings, *e.g.,* by helping them build trusts to share data for common public-sector use cases [4, 9]. Researchers can also draw from the field of *program evaluation* [99] to help cities understand the extent to which already-procured AI, achieves its proposed goals (*e.g.,* of increasing worker productivity, making government services more accessible to residents, preventing car accidents, or other goals).

- Our findings show that many cities, particularly smaller cities, struggle to assess vendors for responsible AI considerations. In today's ecosystem, every city is independently responsible for eliciting relevant information from vendors, and making their own value judgements, effectively on their own. While many cities valued being able to make their own judgements, several participants shared that they wished that there was another centralized body that could conduct reviews to ensure AI vendors met a minimum baseline of ethical behavior. One future direction to explore could include defining such standards and creating a certification system for AI systems, similarly to the existing SOC certification system for security (link).
- More broadly, many participants expressed a wish that private companies be required held to higher ethical standards (*e.g.,* through federal regulation). For example, one participant mentioned that they regularly had to negotiate for minimum data protection standards for residents' data in their interactions with AI vendors. Such rights, the participant argued, could in theory be enacted as a legal expectation for all vendors through federal data privacy legislation. Executive agencies can exercise their rule-making power to establish enforceable standards for commercial AI systems, *e.g.,* so that they comply with existing civil rights and other consumer protection law [24].

Increasing policy attention and resource development has been dedicated towards developing responsible AI initiatives, with the hope that such initiatives can be helpful to some of the highest-stakes AI systems procured by governments. Our study illuminates the complexities of how governments navigate procurement relationships with AI vendors. Our findings point to under-served and important directions for future work that can better support cities in mitigating risks posed by procured AI.

Acknowledgments and Disclosure of Funding

We thank the many city employees who generously shared their time and expertise as participants in our study. We thank the reviewers at the second Regulatable ML Workshop for their feedback that improved our work. This research was supported by a Block Center Responsible AI Seed Grant. HH and NJ acknowledge support from NSF (IIS2040929 and IIS2229881), PwC (through the Digital Transformation and Innovation Center), and the Block Center for Technology and Society at CMU. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of the National Science Foundation and other funding agencies.

References

[1] Procurement: Getting the most out of your demo day, 2017.

[2] Safe or just surveilled?: Tawana petty on the fight against facial recognition surveillance, 2020. Interview.

[3] Tai procurement in a box: Workbook, 2020.

[4] Exploring legal mechanisms for data stewardship, 2021.

[5] What is the difference between rfp and rfb?, 2023.

[6] Advancing governance, innovation, and risk management for agency use of artificial intelligence, 2024.

[7] Census regions and divisions of the united states, 2024.

[8] Govai coalition's open letter to the public, policymakers, and industry., 2024.

[9] The government ai coalition, 2024.

[10] Naspo procurement toolbox: Solicitation methods, 2024.

[11] Request for information: Responsible procurement of artificial intelligence in government, 2024.

[12] Kiana Alikhademi, Emma Drobina, Diandra Prioleau, Brianna Richardson, Duncan Purves, and Juan E. Gilbert. A review of predictive policing from the perspective of fairness. *Artif. Intell. Law*, 30(1):1–17, mar 2022.

[13] Ali Alkhatib and Michael Bernstein. Street-level algorithms: A theory at the gaps between policy and decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery.

[14] Kasun Amarasinghe, Kit T. Rodolfa, Hemank Lamba, and Rayid Ghani. Explainable machine learning for public policy: Use cases, gaps, and research directions. *Data & Policy*, 5, 2023.

[15] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias, 2016.

[16] Aquisition.gov. Part 2 - definitions of words and terms, 2024.

[17] Chloe Autio, Kate Cummings, Brinson S. Elliott, and Beth Simone Noveck. A snapshot of artificial intelligence procurement challenges, 2023.

[18] Johana Bhuiyan. Lapd ended predictive policing programs amid public outcry. a new effort shares many of their flaws, 2021.

[19] Teemu Birkstedt, Matti Minkkinen, Anushree Tandon, and Matti Mäntymäki. Ai governance: themes, knowledge gaps and future agendas. *Internet Res.*, 33:133–167, 2023.

[20] Emily Black, Rakshit Naidu, Rayid Ghani, Kit Rodolfa, Daniel Ho, and Hoda Heidari. Toward operationalizing pipeline-aware ml fairness: A research agenda for developing practical guidelines and tools. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, New York, NY, USA, 2023. Association for Computing Machinery.

[21] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.

[22] Robert Brauneis and Ellen P Goodman. Algorithmic transparency for the smart city. *Yale JL & Tech.*, 20:103, 2018.

[23] Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, and Dylan Hadfield-Menell. Black-box access is insufficient for rigorous ai audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 2254–2272, New York, NY, USA, 2024. Association for Computing Machinery.

[24] Rohit Chopra, Kristen Clarke, Charlotte Burrows, and Lina Khan. Joint statement on enforcement efforts against discrimination and bias in automated systems, 2023.

[25] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 134–148. PMLR, 23–24 Feb 2018.

[26] Cynthia Conti-Cook. A guiding framework to vetting public sector technology vendors.

[27] Catherine Crump. Surveillance policy making by procurement. *Wash. L. Rev.*, 91:1595, 2016.

[28] William Sims Curry. *Contracting for services in state and local government agencies*. Routledge, 2019.

[29] Hannah Quay de la Vallee, Ridhi Shetty, and Elizabeth Laird. Report – the federal government's power of the purse: Enacting procurement policies and practices to support responsible ai use, 2024.

[30] Wesley Hanwen Deng, Boyuan Guo, Alicia Devrio, Hong Shen, Motahhare Eslami, and Kenneth Holstein. Understanding practices, challenges, and opportunities for user-engaged algorithm auditing in industry practice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[31] Wesley Hanwen Deng, Nur Yildirim, Monica Chang, Motahhare Eslami, Kenneth Holstein, and Michael Madaio. Investigating practices and opportunities for cross-functional collaboration around ai fairness in industry practice. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 705–716, New York, NY, USA, 2023. Association for Computing Machinery.

[32] Alameda Police Department. Request for proposals, 2022.

[33] Ravit Dotan, Emmaline Rial, Ana Maria Dimand, and Virginia Dignum. How to manage ai procurement in public administration, 2023.

[34] Julia Edinger. Local governments band together to address use of ai, 2023.

[35] Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

[36] Mihály Fazekas and Jürgen René Blum. Improving public procurement outcomes. *Policy research working paper*, 2(4):2–3, 2021.

[37] Michael Feffer, Anusha Sinha, Zachary C. Lipton, and Hoda Heidari. Red-teaming for generative ai: Silver bullet or security theater?, 2024.

[38] Mary Flanagan, Daniel C. Howe, and Helen Nissenbaum. *Embodying Values in Technology: Theory and Practice*, page 322–353. Cambridge Studies in Philosophy and Public Policy. Cambridge University Press, 2008.

[39] The Organization for Economic Cooperation and Development (OECD). Recommendation of the council on artificial intelligence, 2019.

[40] Center for Inclusive Change. Ai procurement: Essential considerations in contracting, 2023.

[41] Financial Oversight & Management Board for Puerto Rico. Essay: Improving public procurement in pr, 2020.

[42] Aline Shakti Franzke, Iris Muis, and Mirko Tobias Schäfer. Data ethics decision aid (deda): a dialogical framework for ethical inquiry of ai and data projects in the netherlands. *Ethics and Inf. Technol.*, 23(3):551–567, sep 2021.

[43] Marissa Gerchick, Tobi Jegede, Tarak Shah, Ana Gutierrez, Sophie Beiers, Noam Shemtov, Kath Xu, Anjana Samant, and Aaron Horowitz. The devil is in the details: Interrogating values embedded in the allegheny family screening tool. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1292–1310, New York, NY, USA, 2023. Association for Computing Machinery.

[44] Ben Green. The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, 45:105681, 2022.

[45] L Elaine Halchin. Overview of the federal procurement process and resources. Congressional Research Service, Library of Congress, 2006.

[46] MD Romael Haque, Devansh Saxena, Katy Weathington, Joseph Chudzik, and Shion Guha. Are we asking the right questions?: Designing for community stakeholders' interactions with ai in policing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[47] Bernard E. Harcourt. *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*. The Chicago University Press, 2006.

[48] Caroline Haskins. The nypd has misled the public about its use of facial recognition tool clearview ai, 2021.

[49] G Hasselbach, B Kofod Olsen, and Pernille Tranberg. White paper on data ethics in public procurement of ai-based services and solutions, 2020.

[50] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–16, New York, NY, USA, 2019. Association for Computing Machinery.

[51] The Coalition Against Predictive Policing in Pittsburgh. Predictive policing in pittsburgh: A primer, 2020.

[52] Anna Kawakami, Amanda Coston, Hoda Heidari, Kenneth Holstein, and Haiyi Zhu. Studying up public sector ai: How networks of power relations shape agency decisions around ai design and use, 2024.

[53] Anna Kawakami, Amanda Coston, Haiyi Zhu, Hoda Heidari, and Kenneth Holstein. The situate ai guidebook: Co-designing a toolkit to support multi-stakeholder, early-stage deliberations around public sector ai proposals. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[54] Anna Kawakami, Luke Guerdan, Yanghuidi Cheng, Kate Glazko, Matthew Lee, Scott Carter, Nikos Arechiga, Haiyi Zhu, and Kenneth Holstein. Training towards critical use: Learning to situate ai predictions relative to human knowledge. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, page 63–78, New York, NY, USA, 2023. Association for Computing Machinery.

[55] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. Improving human-ai partnerships in child welfare: Understanding worker practices, challenges, and desires for algorithmic decision support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[56] Seyun Kim, Bonnie Fan, Willa Yunqi Yang, Jessie Ramey, Sarah E Fox, Haiyi Zhu, John Zimmerman, and Motahhare Eslami. Public technologies transforming work of the public and the public sector. In *Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work*, volume 31 of *CHIWORK 2024*, page 1–12. ACM, June 2024.

[57] Karen Levy, Kyla E. Chasalow, and Sarah Riley. Algorithms and decision-making in the public sector. *Annual Review of Law and Social Science*, 17(1):309–334, October 2021.

[58] Robert E Lloyd and Clifford P McCue. What is public procurement? definitional problems and implications. In *International Public Procurement Conference Proceedings*, volume 3, pages 2–18, 2004.

[59] Alexandra Sasha Luccioni, Yacine Jernite, and Emma Strubell. Power hungry processing: Watts driving the cost of ai deployment?, 2023.

[60] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. Assessing the fairness of ai systems: Ai practitioners' processes, challenges, and needs for support. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1), apr 2022.

[61] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019.

[62] Sean McGregor. Preventing repeated real world ai failures by cataloging incidents: The ai incident database, 2020.

[63] Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. Assembling accountability: algorithmic impact assessment for the public interest, 2021.

[64] Deirdre K. Mulligan and Kenneth A. Bamberger. Procurement as policy: Administrative process for machine learning. In *Berkeley Technology Law Journal*, 2019.

[65] Deirdre K. Mulligan, Daniel N. Kluttz, and Nitin Kohli. Shaping our tools: Contestability as a means to promote responsible algorithmic decision making in the professions. In *After the Digital Tornado*. Cambridge University Press, 2019.

[66] Matti Mäntymäki, Matti Minkkinen, Teemu Birkstedt, and Mika Viljanen. Defining organizational ai governance. *AI and Ethics*, 2:1–7, 02 2022.

[67] Ankit Fadia Niklas Berglind and Tom Isherwood. The potential value of ai—and how governments could look to capture it, 2022.

[68] Government of Canada. Directive on automated decision-making, 2023.

[69] NYC Mayor's Office of Contract Services. Glossary, 2024.

[70] Gary Peters. Testimony of ritchie eppink: Governing ai through acquisition and procurement, u.s. committee of homeland security, 2023.

[71] Neil Pollock and Robin Williams. Technology choice and its performance: Towards a sociology of software package procurement. *Information and Organization*, 17(3):131–161, 2007.

[72] Eric Prier and Clifford P McCue. The implications of a muddled definition of public procurement. *Journal of Public Procurement*, 9(3/4):326–370, 2009.

[73] EdTech Equity Project. School procurement guide, 2024.

[74] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. The fallacy of ai functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22. ACM, June 2022.

[75] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. Where responsible ai meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), apr 2021.

[76] Christabel Randolph and Marc Rotenberg. The ai red line challenge, 2024.

[77] Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker. Algorithmic impact assessments: A practical framework for public agency, 2018.

[78] Rashida Richardson. Best practices for government procurement of data-driven technologies. *Available at SSRN 3855637*, 2021.

[79] Rashida Richardson et al. Confronting black boxes: A shadow report of the new york city automated decision system task force. *AI Now Institute*, 2019.

[80] Tonya Riley. Feds' spending on facial recognition tech expands, despite privacy concerns, 2022.

[81] Skyler Rispens. San jose, calif., forms nationwide government ai coalition, 2024.

[82] Dorothy E. Roberts. I have studied child protective services for decades. it needs to be abolished., 2022.

[83] David S. Rubenstein. Federal procurement of artificial intelligence: Perils and possibilities, 2020.

[84] David S Rubenstein. Acquiring ethical ai. *Fla. L. Rev.*, 73:747, 2021.

[85] Devansh Saxena, Karla Badillo-Urquiola, Pamela J. Wisniewski, and Shion Guha. A framework of high-stakes algorithmic decision-making for the public sector developed through a case study of child-welfare. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), oct 2021.

[86] Devansh Saxena and Shion Guha. Conducting participatory design to improve algorithms in public services: Lessons and challenges. In *Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '20 Companion, page 383–388, New York, NY, USA, 2020. Association for Computing Machinery.

[87] Matthew Shapanka, Samuel Klein, and Holly Fechner. California establishes working guidance for ai procurement, 2024.

[88] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction, 2023.

[89] Deepa Shivaram. The white house issued new rules on how government can use ai. here's what they do, 2024.

[90] Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M. Kahn, and Adam Perer. Ignore, trust, or negotiate: Understanding clinician acceptance of ai-based treatment recommendations in health care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[91] Mona Sloane, Rumman Chowdhury, John C. Havens, Tomo Lazovich, and Luis Rincon Alba. Ai and procurement - a primer, 2021.

[92] Logan Stapleton, Min Hun Lee, Diana Qing, Marya Wright, Alexandra Chouldechova, Ken Holstein, Zhiwei Steven Wu, and Haiyi Zhu. Imagining new futures beyond predictive systems in child welfare: A qualitative study with impacted stakeholders. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1162–1177, New York, NY, USA, 2022. Association for Computing Machinery.

[93] Harini Suresh, Divya Shanmugam, Tiffany Chen, Annie G Bryan, Alexander D'Amour, John Guttag, and Arvind Satyanarayan. Kaleidoscope: Semantically-grounded, context-specific ml model evaluation. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[94] Elham Tabassi. Artificial intelligence risk management framework (ai rmf 1.0), 2023-01-26 05:01:00 2023.

[95] Ningjing Tang, Jiayin Zhi, Tzu-Sheng Kuo, Calla Kainaroi, Jeremy J. Northup, Kenneth Holstein, Haiyi Zhu, Hoda Heidari, and Hong Shen. Ai failure cards: Understanding and supporting grassroots efforts to mitigate ai failures in homeless services. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 713–732, New York, NY, USA, 2024. Association for Computing Machinery.

[96] Angelina Wang, Sayash Kapoor, Solon Barocas, and Arvind Narayanan. Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy. *ACM J. Responsib. Comput.*, 1(1), March 2024.

[97] Cedric Deslandes Whitney, Teresa Naval, Elizabeth Quepons, Simrandeep Singh, Steven R Rick, and Lilly Irani. Hci tactics for politics from below: Meeting the challenges of smart cities. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[98] Timothy Williams. Facial recognition software moves from overseas wars to local police, 2015.

[99] Amy T. Wilson and Donna M. Mertens. *Program Evaluation Theory and Practice: Second Edition*. Guilford Press, 2012.

[100] Christopher Wilson. Public engagement and ai: A values analysis of national strategies. *Government Information Quarterly*, 39:101652, 01 2022.

[101] Yaniv Yacoby, Ben Green, Christopher L. Griffin Jr., and Finale Doshi-Velez. "if it didn't happen, why would i change my decision?": How judges respond to counterfactual explanations for the public safety assessment. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 10(1):219–230, Oct. 2022.

[102] Tom Zick, Mason Kortz, David Eaves, and Finale Doshi-Velez. Ai procurement checklists: Revisiting implementation in the age of ai governance, 2024.

A Related Scholarship

We group related scholarship on responsible governance of public AI into two categories: (1) scholarship motivating why responsible AI is particularly important to study within local government contexts; and (2) scholarship specifically targeting procurement as an intervention point for responsible AI.

A.1 AI in the Public Sector

Scholarship has documented a recent surge in public sector adoption of AI [77, 56, 25, 35, 97, 52, 57]. At times, the subsequent harms of these AI systems has further complicated government's relationship with marginalized communities, such as child protection services (CPS) [35, 82, 92] or local police [47, 51, 18]. For example, past work has demonstrated how public AI algorithms trained on biased human decisions also replicate historic biases in deployment, such as disproportionately flagging Black defendants as high-risk of recidivism [15] or targeting poor families for CPS investigations [35]. In these and other high-stakes domains where AI "displaces discretion previously exercised by policymakers" [64], scholars have argued that public AI represents a fundamental shift in how public *policy* is formulated and implemented [13, 43]. Thus, a growing number of both technical and policy interventions have emerged to facilitate more responsible development and governance of public-sector AI [63, 77, 53, 42, 12, 100, 44], organized around principles, such as fairness, transparency, accountability to the public, and democratic participation. More broadly, research has pointed to the importance of *organizational AI governance* frameworks (*e.g.,* the NIST AI Risk Management Framework [94]) to support organizations in establishing consistent standards and clear lines of accountability for AI systems [66, 19].

Although an abundance of responsible AI resources exist, organizations often fall short in implementing best practices. To better understand this disconnect, a emerging body of empirical HCI research aims to understand the experiences and needs of practitioners that oversee or interact with AI systems, on-the-ground [50, 31, 101, 55, 90, 52]. Many of these studies highlight how proposed responsible AI interventions are disconnected from practitioners' actual needs, which are often shaped by organizational culture, pressures, and incentives [50, 31, 52, 60, 86, 85, 55, 52]. We adopt this methodological approach to study the public sector, specifically public procurement of AI. Our study contributes to the emerging body of empirical studies that examine public sector employees' use of AI technologies [86, 85, 55, 52, 17].

To our knowledge, our study is the first systematic investigations of city employees' AI *procurement practices* using qualitative methods. Despite widespread attention to AI procurement, to our knowledge the only prior work that spoke with government employees about procurement and AI is [17]. Their study differs from ours in that they (1) primarily spoke to federal officials, and (2) adopted a "bench research" methodology to identify major themes.

A.2 Public Procurement for Responsible AI

In response to increasing incidents of harm caused by public AI [62], experts have called for governments to adapt their existing procurement processes to the unique challenges and risks posed by AI [83]. Several groups have developed *practical guidance* and *readily-adoptable resources* centered around AI procurement [78, 49, 73, 68]. These resources, intended to be used by government employees, include evolving guidelines, regulation, tools, vendor repositories, and templates to guide public sector procurement practices.[7] Many of these resources are targeted to the steps of a conventional procurement process, *e.g.,* items to add to an RFP [3, 26, 33, 84, 22, 73], guidance on how to score AI proposals [79], and AI-specific terms for procurement contracts [32, 40]. While a handful of these resources were empirically validated [79], we still lack a broader understanding of if, and how, governments have made use of these resources. Thus, we examine if existing resources address city employees' primary needs for support.

Moreover, several scholars have pointed out how the public procurement process poses opportunities to encourage better responsible AI practices. Individual cities can use their purchasing power to walk away from harmful AI systems and negotiate on behalf of residents' best interests, *e.g.,* requiring that vendors implement harm mitigation steps [26]. Furthermore, some argue that incorporating

[7] We refer the reader to [33] for a more detailed review of existing resources.

responsible AI considerations into purchasing decisions can result in broader market shifts that incentivise best practices, especially for technologies that are exclusively sold to governments, *e.g.,* policing technologies [36, 27].

More broadly, our research coincides with a landmark year of government action focused on improving AI procurement practices. The U.S. Federal government's "AI M-Memo" [6, 89] is perhaps the most comprehensive action taken by the federal government to date, and also initiated a broader conversation amongst key stakeholders about the role of procurement in responsible AI by soliciting public comments [11, 102, 29]. Several U.S. state and local governments have followed suit in adopting their first AI procurement guidelines [87].

Another key action that occurred as we were conducting interviews was the formation and announcement of the Government AI ("GovAI") Coalition [9, 34, 81], a group composed of over 1,000 members representing 350 participating U.S. governments, founded to "give local governments a voice in shaping the future of AI". Participating cities are encouraged to adopt shared AI governance practices based on resource templates created by coalition members. The coalition envisions that by working together cities can ask more of AI vendors: *e.g.,* that they share basic information about their AI systems with cities [8]. Thus, we believe our research is of timely importance to inform evolving policy efforts and their implementation.

B Defining "procurement"

Despite the differences, all the definitions of "public procurement" encountered share fundamental similarities: they all concern the process of bringing in goods and services that are developed externally, to achieve the goals of a public sector entity. They differ, however, in specific components of the process. For example, while the United States' federal definition emphasizes a "competitive" purchasing process, denoting the exchange of money as part of procurement, some local governments, like New York City, have definitions that are broader, encompassing all functions related to obtaining goods and services whether or not money changes hands [16, 69].

In this paper we do not adopt a single definition of public procurement as methodologically we chose to leave such distinctions to our interviewees who were encouraged to discuss whatever processes and components they personally and professionally associated with public procurement. Given the broader diversity of the term, as would be expected, we observed differences across municipalities in what types of acquisitions and activities participants deemed as falling under the umbrella of "procurement". For example, procurement departments often did not oversee governments' acquisition and use of free technologies.

C Interview Protocol

We began the interview by reminding the participant of our informed consent protocol (approved by our institution's IRB board), and asking for their consent to record.

Introduction The goal of this interview is to learn more about existing procurement practices specifically for artificial intelligence, or AI, technologies in your city. We adopt a wide definition of AI as "any machine-based system that can make predictions, recommendations, or decisions". This would include technologies such as facial recognition, gunshot recognition technology, resume screening technology, ChatGPT, etc.

Our goal is not to assess your practices, but rather to identify needs and opportunities for researchers as partners to support US cities.

Q1.1: Can you tell me a bit about your current role, and any past work experiences or responsibilities relating to artificial intelligence?

Q1.2: Have you ever been involved in a past procurement of an artificial intelligence technology?

- If YES: How were you involved?
- If NO: Has your [agency] ever considered or talked about procuring AI?

Walk-through The goal is to understand how a "typical" AI procurement occurs in the city. Our goal is not to impose structure on the participant's description, but rather allow them to describe how they personally view/understand the procurement process.

If it doesn't come up naturally, we can prompt them to reflect on specific parts of procurement, such as (1) Planning, (2) RFP writing, (3) Evaluating Vendors, (4) Contracting, (5) Designing/Building/Evaluating the AI, and (6) Deployment, and (7) Post-deployment.

Q2.1., Walkthrough. Can you briefly walk us through how a typical procurement involving an artificial intelligence technology would occur in your city? We're specifically interested in understanding any difference between a standard technology procurement, vs. a procurement involving AI.

- If never procured AI: *e.g.,* imagine your city is considering procuring an enterprise-level generative AI product, like a chatbot to screen 311 questions.

Drill-down prompts on specific parts of the procurement process:

Planning (Problem Formulation):

1. What does your city do to plan for the procurement before the RFP (request for proposal) writing stage?
2. (if not covered) Pre-RFP, how does the agency identify that an AI tool might be a part of the solution (rather than a tool that does not use AI)?
3. (if not covered) Do you have a process for evaluating the risks of a proposed AI technology before RFP writing?
   - If YES: What about potential mitigation processes for these risks?

RFPs:

1. Is there anything different in the content of the RFP for AI procurements, compared to standard technology procurements that do not involve AI?
2. (if not covered) In the RFP, do you ask vendors questions about potential risks and mitigation strategies?

Evaluating Proposals:

1. How does your city evaluate proposed AI solutions? We are especially interested in differences between evaluating standard technology vs. AI proposals.
2. (if not covered) What information do you ask vendors to report in their proposal? Do you ever encounter "trade secrecy" claims?
3. (if not covered) What measures do you expect them to report? Do they validate that the technology works as claimed using data from your city?

Contracting:

1. Are there any differences in the contracting process for AI vs. non-AI (standard technology) technologies?
2. Are there specific terms and conditions that you include in AI contracts?
3. Can you share a past contract for an AI technology with us?

AI Design, Development, and Evaluation:

1. How are people from your city involved with the design, development and evaluation of AI technologies under contract?
   - If YES: How were you involved? What type of feedback did you give?
2. How often do vendors make changes to their technologies (like updating or improving it using data from your city) before they are deployed?
3. How do vendors evaluate the AI solution they have designed and developed to make sure it fits your use case?

- Do they use data from your municipality for evaluation?
- What kind of measures do they look at and report to you?

AI Deployment:

1. How often do vendors (or the city) provide training or onboarding for people who will be using the AI?
2. How are agency workers involved in deciding the way the AI is used in their everyday practice?

Post-deployment:

1. How do you oversee and monitor deployed AI technologies?
   - What is the vendors' responsibility?
   - What if something goes wrong? (liability)

Q2.3 (if unclear) Can you remind me of who in your city is involved or oversees each phase of this procurement process?

Q2.4 (if unclear): Do you believe the process that we just went through together is representative of most AI procurements in your city (if relevant: beyond that specific example)?

Q2.5: Are there any existing policies in place that target the procurement of AI technologies specifically?

- If YES:
  - Can you share your city's policies/guidelines with us?
  - How long have these policies been in place?
- If NO:
  - Is this something you anticipate being developed in the near future, or something that has been discussed?

Q2.6: Can you direct us to your city's general procurement policies that may be applicable to AI technologies? e.g., such as data privacy policies?

Q2.7: Are there any AI technologies that come to be used through processes outside of the traditional procurement process? (e.g., research partnerships, foundations, donations, or free tools?)

- Do these technologies undergo a similar "vetting" process to procured technologies?
  - Do similar people evaluate these proposals?
  - Do similar people oversee or monitor their deployment?

Q2.8: Does your city consider opportunities to engage with residents who may be affected by an AI tool during the procurement process?

Challenges & Desires The goal is to understand the participant's needs and desires to improve the procurement process.

For the last part of our interview, we'd like to understand your opinions and wishes for improving AI procurement.

Q3.1: What do you believe are the main challenges or "pain points" for AI procurement in your city?

- Do you have any suggestions as to how cities could improve their procurement of AI?
- (if relevant) Do you have any examples where [this challenge] happened in the past?

Q3.2: Can you imagine any new resources that could help you address these challenges?

- What resource format would be most helpful? ex: Checklists? Templates? Trainings?

D Cities' AI Use Cases

Table 2 groups examples of AI adopted by municipalities into five categories based on their intended usage. In our discussions, employees in each city shared at least one example that they were aware of belonging to one of these five categories.

Interestingly, not all of the employees that we interviewed were aware that other employees in their city had already procured or adopted AI technologies: for example, one city employee stated that to their knowledge, their city "has never purchased anything AI related", whereas their colleagues stated that the city in fact has.

Type of AI technology	Examples
Facilitating resident communication	Translation services, chatbots, 311 assistance, public meeting summaries
Law enforcement	License plate readers, gunshot detection, object detection
Smart cities/urban planning	Sensors to track service utilization, accident tracking, snow plow routing
Assisting bureaucratic decision-making	Funding allocation, service allocation, school bus routing
Workplace productivity tools	Chatbots, image generation, voice generation, coding assistants

Table 2: We grouped the AI systems that municipal employees discussed procuring or adopting in interviews, into 5 categories based on their intended usage. We provide anonymized examples of types of AI systems that were mentioned in each category. Employees in each city shared at least one example that they were aware of belonging to one of these five categories.

E Limitations

We acknowledge several methodological limitations of our study, many of which pose directions for future work. We leveraged our personal networks and snowball sampling to recruit participating cities and employees for our study. As such, our sample was skewed towards large cities who had leaders that were already interested in and knowledgeable about artificial intelligence. Similarly, our recruitment criteria may have led us to recruit participants who stood at the "forefront" of their citys' emerging AI practices, a methodological limitation shared by related empirical studies on topics related to responsible AI [30, 50, 75]. To address these limitations, we intentionally tried to recruit smaller cities and participant roles that were under-represented in our sample. While our focus on U.S. cities enabled us to draw productive comparisons across jurisdictions, we believe that understanding generalizability and distinctions across governments in other countries is an important direction for future work. Finally, we acknowledge that our decision to focus on city employees as a key stakeholder group in public-sector AI governance neglects the perspectives of other important actors, such as AI vendors, civil society and other members of the public, and impacted communities.

F Extended Descriptions of Cities' AI Procurement Review Processes

Several cities that we spoke with had already introduced specific changes to their existing procurement practices for AI, beginning in 2021 onwards. Notably, many participants felt that it was "early days" in revising their AI review processes: for example, participants were in the midst of overseeing their first formal AI procurement, conducting their first AI risk assessments, and revising their processes more broadly. With this rapidly evolving landscape in mind, we group changes cities had made so far to their practices into 5 categories, based on their goals. We discuss each of the interventions in detail below, providing examples when appropriate.

Vendor reporting requirements Several cities instituted additional reporting requirements to ask vendors for important information about AI systems. Cities could mandate that vendors complete the reporting requirements by adding them as required items on an RFP or solicitation, or asking a vendor to provide them separately for purchases without a solicitation. Participants believed that learning additional information about an AI system could help cities with making more informed purchasing decisions, risk assessments, and contract negotiations.

When deciding what to ask of vendors, several participants shared that their city started with the list of questions from Government AI Coalition's "AI FactSheet" [9], a resource designed to support local governments in understanding third-party AI systems. The factsheet asks vendors to report "essential technical details" such as on what data the AI was trained, under what conditions the system was tested, the values of relevant performance metrics, and measures taken to promote values of fairness, robustness, and explainability.

AI risk or impact assessments Participants conducted additional risk or impact assessments to better understand the possible positive or negative impacts of procured AI systems. While some cities conducted such assessments in an informal or ad-hoc way, others had started to standardize assessment processes by creating assessment templates with lists of questions and considerations. Different cities also conducted risk or impact assessments at different phases of the procurement process: some assessment instruments could be completed based on a "purpose statement" for AI, before a specific vendor or AI system is identified. In contrast, other risk assessments can only be completed once a concrete system has been identified, e.g., they require knowledge of the system's performance.

The role and purpose of these assessments varied across interviewed cities. In many cities, the risk assessment had no immediate outcome, but employees were encouraged to take action to manage and if possible, mitigate potential risks identified in the process. Beyond informing mitigation steps, some participants also used risk assessments to triage AI solutions into "high" or "low" risk categories, which then determined subsequent requirements for review and oversight. For example, one city required high-risk AI to have additional reporting requirements, further risk assessment, usage protocols, and regular post-deployment monitoring. Participants viewed risk triaging as a way to reduce reviewing burden and better allocate their limited technical expertise. One participant who conducted AI risk assessments explained:

> "[When triaging risk], we're just trying to get a sense of how thorough a review we need to do, because we're working with very limited capacity and resources. So we've got to decide: is this a low-risk system that we can just do a really quick look at? Or is this going to be something really sensitive and safety-impacting, rights-impacting, that we need to dedicate a lot of our time to?" (P10)

Participants also noted how risk triaging was also time-saving for their colleagues on the other end trying to purchase the AI, as put by one employee: "If it's low risk, I'll approve it, and you'll be on your way tomorrow!"

Independent evaluations While less common, some participants conducted their own independent evaluations (or "audits") of third-party AI systems. Participants were motivated to conduct their own evaluations for a variety of reasons, such as wanting to measure constructs that vendors had not reported, or calculate these measures using their own data. For example, one participant described an experience where they wanted to understand how a translation model's accuracy rates differed across languages, but the vendor didn't give the city "any meaningful information" about the system's

performance. When the employee realized they had API access to the model, they decided to "go into the platform [themselves]" to calculate performance metrics.

Another participant was motivated to design and conduct their own independent evaluations because they believed demos to "show them [the AI] works" were insufficient to evaluating AI, "because [vendors] give a demo of the one thing that works". The participant appreciated their city's ability to define what measures were most important for them:

> "Early on, doing early experimentation very cheaply and fastly helps weed out a lot of things. [...] The people that will best know what might be helpful, are the employees of the city. It's not like some CIO in the clouds coming like: 'Oh, I have determined that this would be helpful'." (P18)

AI contracting terms Procurement contracts specify legally enforceable obligations for both cities and vendors, such as the agreed price, statement of work, the vendor's support responsibilities, and an outline of how disputes will be resolved. Participants across cities pointed to the importance of including clear expectations of vendors in the contract, as put by one participant (P17): "The problem with holding a vendor accountable, was you got to make sure you actually have somewhere where it's documented what we're going to hold them accountable to".

To hold AI vendors accountable, some cities created contract language templates that spelled out expected risk mitigation practices for procurements involving AI. Example contract terms included requirements for vendors to regularly monitor system performance, train city users on how to operate the AI, respond in a timely manner to incidents where AI causes harm, and comply with data privacy legislation. Participants who were members of the Government AI Coalition shared their intentions to adopt terms from the GovAI's Vendor Agreement contract template, believing that cities could ask for more from vendors when they "stand together" and adopt similar terms.

Prototype deployments Typical procurement contracts require significant commitment from a city, both financially and legally through a binding contract. As an alternative, several participants instead preferred to "try out" emerging AI technologies via fixed-length contracts. One department leader who ran a prototyping program explained how they "work within the constraints of state procurement law" by "paying for short-term, small-scale prototypes on the order of weeks to months, that are under requirements for payment thresholds". Successful prototypes can then go through a formal solicitation after the short-term contract has ended.

Beyond helping understand if an AI technology is financially worthwhile, the participant believed that prototype contracts can help surface potential risks posed by procured AI. The participant described an example of how they deployed a prototype of an AI chatbot that "started hallucinating and interacting with people in really unexpected ways", which "sparked an interesting conversation in our community about the risks of AI tools". When reflecting on the experience, the participant was grateful that they had procured the service using a short-term contract:

> "This is exactly why we need a program like ours, to 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 just deploying [AI] than by trying to plan out every detail in advance." (P2)

F.1 Practice vs. policy?

In many cities, employees made one or more of the above changes to their procurement practices simply by adjusting their existing practices, *e.g.,* by electing to include vendor reporting requirements in an AI RFP. Some cities decided to make these changes in their practices more formal or mandatory for vendors or city employees, by adopting policies or passing laws that required them. For example, one department leader walked through how their city's formal AI policy spelled out mandatory steps, such as a risk assessment, that city employees must complete for any AI procurement. The participant viewed the policy, which was passed by their city council, as an "accountability trigger" to incentivise compliance for both colleagues and vendors:

> "Council adopted the policy. So you can't just say no. I'm going to have some leverage to say, we can't just say we're not going to do this. [...] [The policy] is

> really meant to be a way to say the city is going to be taking this on, these are our values." (P13)

Participants in another city shared that while ideally someday they would like to institutionalize their practices via a formal policy, at the time of interviewing, they did not yet have one:

> "We very intentionally have not put out a [formal] AI policy yet, because we wanted more [community and government] input on it. And the space, especially in 2023, was very new for us. So we wanted to get a better understanding before asking our leadership to pass a policy." (P7)

This city has since adopted a formal AI policy following engagement with the community, experts, and agency staff.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We accurately describe our research questions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We include an extended discussion of our study's limitations in Appendix E.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification:

   Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We include our interview protocol in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We unfortunately cannot share interview transcript data due to our informed consent conditions and to preserve participant anonymity.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.

- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification:

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: The protocol is in Appendix C.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: The study was reviewed and approved by our university IRB.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.