Studying Up Public Sector AI: How Networks of Power Relations Shape Agency Decisions Around AI Design and Use

ANNA KAWAKAMI, Carnegie Mellon University, USA
AMANDA COSTON, Carnegie Mellon University, USA
HODA HEIDARI*, Carnegie Mellon University, USA
KENNETH HOLSTEIN*, Carnegie Mellon University, USA
HAIYI ZHU*, Carnegie Mellon University, USA

As public sector agencies rapidly introduce new AI tools in high-stakes domains like social services, it becomes critical to understand *how* decisions to adopt these tools are made in practice. We borrow from the anthropological practice to "study up" those in positions of power, and reorient our study of public sector AI around those who have the power and responsibility to make decisions about the role that AI tools will play in their agency. Through semi-structured interviews and design activities with 16 agency decision-makers, we examine how decisions about AI design and adoption are influenced by their *interactions with* and *assumptions about* stakeholders situated *within* these agencies (e.g., frontline workers and agency leaders), as well as those *above* (legal systems and contracted companies), and *below* (impacted communities). By centering these networks of power relations public sector agencies are embedded in, we clarify how a range of infrastructural, legal, and social factors–including impoverished understandings of frontline workers' concerns, constant pressures from other institutions of high power, and uncertainty around how to connect and communicate with impacted community members–collectively disincentivize agencies from making decisions about AI design and adoption that are informed by stakeholders outside of their immediate leadership and development teams. Agency decision-makers desired more practical support on operationalizing value-sensitive and participatory design approaches, to help overcome the power and knowledge differentials they perceived between them and other stakeholders (e.g., impacted community members). Building on these findings, we discuss implications for advancing a power-conscious research and policy agenda for public sector AI.

Additional Key Words and Phrases: Public sector AI, Studying up, Semi-structured interviews, AI procurement

ACM Reference Format:
Anna Kawakami, Amanda Coston, Hoda Heidari, Kenneth Holstein, and Haiyi Zhu. 2023. Studying Up Public Sector AI: How Networks of Power Relations Shape Agency Decisions Around AI Design and Use. In . ACM, New York, NY, USA, 24 pages.

1 INTRODUCTION

Public sector agencies across the United States are rapidly deploying AI-based tools to assist or automate complex, high-stakes decisions traditionally made by frontline workers—for example, in social services, public health, criminal justice, and education [3, 28, 52, 56]. Agencies have been motivated to explore the use of these new AI tools to help overcome resource constraints and limitations in human decision-making (e.g., biases in worker decisions) [1, 9, 49]. However, the use of AI tools in these domains have been met with significant contention. A growing body of research and public media concerns the bias [9, 33], validity [10], and lack of transparency [6, 42, 59], amongst other concerns, arising from AI tools deployed in the public sector. To address these concerns, the human-computer interaction (HCI) and machine learning (ML) communities are increasingly exploring approaches to improve the design and use of public sector AI tools. For example, prior work on AI-based decision support tools has studied racial disparities in algorithmic predictions and ways to mitigate them [7, 18, 33, 51], interface-level interventions for frontline workers using

*Co-senior authors contributed equally to this research.

these tools [19, 22, 23, 25, 55], and community members' concerns and desires around them [6]. This work has substantially advanced our understanding of the social impacts of public sector AI systems, and opportunities for improving how they are designed and used.

The challenges outlined by prior work point to a crucial, yet under-examined question: How do public sector agencies make decisions about *whether* to create or adopt a new AI tool in the first place? While prior theoretical work has established the consequential nature of decisions and assumptions made in early stages of model design (e.g., in problem formulation [35, 53]), we still lack an empirical understanding of *how* these decisions are made in practice. Addressing this question requires us to redirect our focus to a small subset of public sector agency stakeholders who hold significant amounts of institutional power over decisions on AI use and design. In doing so, we borrow from anthropologist Laura Nader's 1972 call to "*study up*" – reorienting anthropological fields of study around those who have power and responsibility to shape social systems and norms [32]. In the study of AI systems, recent work has urged academic researchers to reorient the studies of algorithmic fairness [2] and machine learning datasets [30] around issues of power, demonstrating the value of studying up technology as a means to uncover broader social and infrastructural challenges impeding the responsible development and use of AI systems. More broadly, literature in the Science and Technology Studies have contributed a range of concepts and frameworks to understand relationships between power and technology (e.g., in [11, 26]). We extend these prior lines of work, by grounding an empirical understanding of how power structures impact on-the-ground decisions around public sector AI–a topic that has thus far mostly been discussed theoretically.

In this paper, we *study up* public sector AI via the perspectives and experiences of agency decision-makers: those typically occupying upper levels of a public sector agency's organizational hierarchy, holding the power and responsibility to progress, halt, or otherwise shape whether and how AI tools are designed and used. We focus on public sector AI in social services, where "social services" refers to a broad range of government services intended to benefit the community, such as housing, education, child welfare, and healthcare. Much of our research-based knowledge on AI for social services are from the perspectives and experiences of those directly impacted by higher-up decisions (e.g., frontline workers [24, 43], community members [6, 47]). To better understand the role of agency decision-makers, we ask the following research questions:

- What perceptions, practices, and challenges shape agency decisions to move forward with the creation or use of AI tools?
- What opportunities exist to improve decisions around the creation or use of AI tools?

To address these questions, we conduct semi-structured interviews and design activities with 16 agency decision-makers across three public sector agencies in the United States. In our analysis, we most prominently observed ways in which agency decision-makers' *interactions with and assumptions about other stakeholders* shaped their decisions around AI. To contextualize agencies' decisions about AI in the broader web of social and infrastructural systems they reside in, we organized our findings around the power relations public sector agencies share with sectors and stakeholders situated *within*, *above*, and *below* the agency. Our findings surface how centering this network of power relations clarifies the different factors shaping agency decisions around AI design and use:

- **Misconstrued concerns around the value of AI tools within agencies:** Prior academic and grey literature has often critiqued the outcomes of agency decisions, depicting the public sector agency as a monolithic entity united in their decisions to adopt or drop AI tools (e.g., [41]). Our findings surface how existing depictions of agencies overlook the inter-agency contention that exists amongst workers around these decisions. Participants described how

agency workers–including those with relatively higher institutional power (e.g., agency leaders, scientists) and frontline workers–held dissenting perspectives on the validity and value of AI tools. While these workers with higher institutional power advocated for increased community engagement in their agency and more reflexive deliberation on technical design decisions, their perspectives were sometimes ignored or questioned by their colleagues, the majority of whom were perceived to be interested in advancing their agency's use of AI tools. Agency leaders and developers promoting the value of new AI tools described efforts to mitigate frontline workers' concerns towards AI, which they believed to be especially pervasive and persistent, by "convincing" frontline workers of the value of AI tools. Talking with frontline workers who were involved in piloting and providing feedback on new AI tools, we observed disconnects between the reasons underlying frontline workers' concerns and the reasons often attributed to them by agency leaders and developers. (Section 4.1)

- **Constant pressures from other institutions of high power:** Our findings suggest that agency leaders and creators of public sector AI tools experience constant pressures from other institutions of high power (e.g., legal systems and state courts, private companies), that shape their AI design and use decisions. Whereas prior empirical research has tended to focus on understanding factors within the organization as its unit of analysis (e.g., organizational governance, worker interaction, or model design factors), our findings help contextualize how agencies' interactions with external institutions of high power may impact those organizational practices and decisions. For example, participants felt frustrated with their inability to interrogate ethical considerations when adopting AI tools from private companies. They felt constrained by unfavorable working relationships predefined by procurement contracts, which disincentivized private entities from being transparent in their model development and evaluation process. Other participants shared concerns that their AI tools, while intended to be used by frontline workers, could be misused as evidence for or against the agency in legal court cases. Participants described that these legal pressures were further exacerbated by constrained communication channels, including the lack of interactions with legal experts when creating or adopting AI tools. (Section 4.2)
- **Disincentives to hearing and empowering impacted communities:** While academic research has proposed a multitude of approaches (e.g., [36, 46]) to expand stakeholder involvement in AI design and development, most participants perceived infrastructural and communication barriers between them and impacted community members external to their agency; these barriers created additional disincentives for connecting with or involving impacted communities in their AI design process. Participants–who were experts in leading agencies or developing technology but had minimal to no experience in community outreach–felt they were ill-equipped to make decisions about how to involve the community: which groups to reach out to, how to develop sustainable relationships with them, and how to mitigate power imbalances when involving them. As a result, participants often described that interactions with the communities they serve were either very limited, heavily constrained, or non-existent. (Section 4.3)

By studying up the network of power relations public sector agencies operate within, we advance an empirical understanding of why it may be challenging to actualize design, modelling, or policy improvements in practice, even when proposed as best practices in the research literature. While agency decision-makers hold direct responsibility for making the decisions that legitimize existing practices, our findings surface a range of legal, infrastructural, and social barriers to responsible design that agencies are currently ill-equipped to address alone. Developing the systemic conditions and practical resources needed to shift towards an improved practice requires urgent attention

from the research, design, and policy communities. We conclude in Section 5 with implications for advancing a power-conscious agenda for research and policymaking around public sector AI.

2 BACKGROUND AND RELATED WORK

2.1 AI in Public Sector Social Services

As public sector agencies aim to mitigate resource constraints and improve the decision-making quality of their staff, several have turned to exploring the use of new AI tools. In social services alone, AI tools have been rapidly deployed to assist decisions around child maltreatment screening [9], homeless services [27], education [20, 40], and predictive policing [45]. However, the deployment of AI tools in social services, and especially predictive AI tools, has also introduced a myriad of ethical and social concerns. For example, researchers have raised questions around systemic biases embedded in AI tools [7, 14], the validity of the models used [10, 37], and the (lack of) support for frontline workers using them [24].

Towards mitigating the potential social harms of AI tools introduced to the public sector, prior work has proposed various approaches to improve technical design decisions. Acknowledging the value-based tensions across the many stakeholders of AI tools [22], many of these approaches aim to incorporate stakeholder-specific values and knowledge into the design of algorithms, for example, through expanding participation along the AI design process. Prior work has proposed a myriad of participatory approaches [12] (e.g., value-sensitive algorithm design [58], deliberation-based participatory algorithm design [57]) to solicit and operationalize stakeholders' desires and values into the design of algorithmic systems. Such work has often also surfaced the challenges of actually incorporating multiple, conflicting values into an algorithm. For example, Møller et. al. [22] discusses deviations in what notions of value that developers and caseworkers believed should be considered in metrics for algorithmic decision-making systems.

A growing body of literature has shed light on how AI tools are actually integrated into social service agencies (e.g., [24, 43]), implicating how design approaches and interventions proposed in the academic literature may, in the future, *actually* fit into real-world contexts. This work has begun to surface additional challenges arising from sources beyond the model–for example, from organizational pressures and incentives, or power imbalances across stakeholders [24]. These findings raise questions around how organizational and social factors may shape practices and norms surrounding the design and deployment of public sector AI. This work has largely relied on the insights of AI developers who build and frontline workers who use AI tools, leaving open the question how those holding positions of higher power and responsibility, like directors and managers in public sector agencies, influence these decisions. Our work aims to illuminate the perspectives and experiences of those holding positions of higher power who ultimately drive decisions around the design and deployment of AI.

2.2 Studying Up Algorithmic Systems

Towards better understanding the broader systems of power and privilege shaping the impacts of algorithmic systems, a growing chorus of researchers have called for a reorientation of the study of AI around issues of *power* [2, 15, 30, 44]. Prior work has examined the downstream consequences of decisions made by those who develop and govern AI technologies, to examine tensions between impacted stakeholders and institutions of power (e.g., [13]).

Drawing on anthropological practices to study those in higher positions of power, researchers have increasingly recognized that looking downwards only provides a partial understanding of the broader social and structural mechanisms that shape AI systems. Recent work has called for more researchers to shift their gaze upwards, to study the practices of more powerful stakeholders. These

calls have reflected on the relevance of Laura Nader's 1972 proposal to "*study up*" [32], urging anthropologists to reorient their studies to examine how the practices of powerful institutions and authority shape everyday experiences and norms. In recent years, Barabas et. al. [2] discussed the need for a similar reorientation around power in the study of data science, explaining: "The political and social impacts of algorithmic systems cannot be fully understood unless they are conceptualized within larger institutional contexts and systems of oppression and control." Miceli et. al. [30] further extend these calls to re-examine machine learning data issues typically labeled as problems of "bias," demonstrating how viewing them as problems of power surfaces often overlooked factors around social contexts (e.g., labor conditions, epistemological stances) underlying data problems.

To holistically understand the social mechanisms shaping *how* AI systems could bring downstream harm to vulnerable communities, we extend this line of work, revisiting questions around the design and deployment of AI systems in the public sector. We shift our gaze upwards to focus on the decisions and assumptions made by stakeholders with relatively high amounts of decision-making power in public sector agencies, and the broader social systems and infrastructures surrounding them that shape their decisions around AI.

3 METHODS

To "study up" the perspectives of those in positions of power to shape the design and deployment of public sector AI, we conducted semi-structured interviews and design activities with 16 participants across three public sector agencies. We recruited agency stakeholders who have experience making or shaping decisions around the design and deployment of public sector AI technologies. In the following, we describe our recruitment process, study procedure, and analysis method.

3.1 Public Sector Agencies and Participants

To understand perceptions, practices, and challenges shaping the design and deployment of public sector AI across contexts, we aimed to recruit participants from a range of public sector agencies across the United States. Our team searched for U.S.-based state, city, or county-level public sector agencies with human service departments (e.g., child welfare, predictive policing) that have previously or are currently considering the use of AI tools, had previously deployed then stopped using AI tools, and/or are currently using AI tools. We found 19 agencies fitting these criteria based on information available on public agency websites or news articles. At each of these agencies, we found contacts in leadership positions (e.g., directors of departments), and emailed them stating our research goals and interest in starting a conversation. Out of the 19 agencies, contacts from four of the agencies responded and agreed to having an initial conversation. Out of those four agencies, three agencies agreed to participate in our study.

From these three agencies, we aimed to recruit participants in occupations that are typically tasked with making or informing decisions around the use or design of AI tools. On aggregate, we interviewed directors of human service departments, directors and managers of teams responsible for developing AI tools, and researchers and analysts who are involved in building or analyzing AI tools. In one agency, our contact connected us to frontline staff who played a role in piloting AI tools or had otherwise also discussed the use of AI tools with developers and leaders in their agency. In our communication to each agency, we requested study participation from the same set of occupations. The actual occupations and numbers of participants involved in our study were determined by who our contacts at the agencies connected us to.

We refer to participant occupations using four stakeholder groups: **1**)*Agency Leaders* (L): Individuals in director or managerial roles who typically are involved in making agency- or department-level decisions, **2**) *Research and Development Workers* (RD): Individuals in research, development, and/or

analysis teams internal to a given public sector agency who are involved in the creation or evaluation of AI tools used by the agency, **3**) *Frontline Leaders* (FL): Individuals who work in leadership roles within human service departments in a public sector agency and who typically are in closer communication with frontline workers than Agency leaders, and **4**) *Frontline Workers* (FW): Individuals at a public sector agency whose occupations bring them in direct contact with the community their agency serves. Our study includes six Agency Leaders, six Research and Development Workers, two Frontline Leaders, and two Frontline Workers.

3.2 Semi-Structured Interviews and Design Activity

We conducted 90 minute study sessions with each session including a semi-structured interview and design activity. The semi-structured interview included three sections. First, we asked participants about their background, for example, questions related to their current role in the public sector agency. Next, we asked participants about their current and past practices for making decisions around whether to design, deploy, or adopt a given AI tool in their public sector agency. This involved asking questions about what considerations aided these decisions (e.g., goals for introducing algorithm, empirical evaluations conducted). In the next section, we probed more deeply on challenges they faced in past experiences designing, deploying, or adopting AI tools.

To complement our semi-structured interview, we also conducted a design activity. Through the design activity, we invited participants to ideate future considerations that they believed should aid decisions around the design, deployment, or adoption of public sector AI tools. We asked follow-up questions to better understand participants' past experiences and challenges that shaped their ideas. By prompting participants to ideate considerations for a hypothetical future state, we opened opportunities for participants to also reflect on and identify a broader set of challenges they currently face that otherwise may not have surfaced during the semi-structured interview. For example, participants who described a future need to expand community participation in the AI design process also reflected on current barriers their agency faces to supporting external participation in their design process. In this case, because community participation does not currently exist at the participants' agency, participants seldom brought up these barriers during the semi-structured interview, which was more focused on understanding current practices and challenges. To support the design activity, we shared our screen and shared a link to an online board on Figma, a collaborative web application that accommodates multiple users. As participants ideated future considerations they believed should aid the hypothetical design and deployment of AI technologies, we used online post-its to help document and organize their ideas.

All participants completed the design activity. Participants who were not involved in decisions around whether to design and deploy AI tools (i.e., frontline managers and frontline workers) moved onto the design activity after completing the first portion of the semi-structured interview (on participant background). Participants who participated in all three sections of the semi-structured interview completed the design activity immediately afterwards. All 16 participants consented to having their audio recorded, shared screen recorded, and having notes from the design activity recorded and saved. We requested individual study participation from participants. Some participants preferred to participate in the research study in pairs because they felt that the pair would provide a more extensive set of insights together, or because they wanted to provide the other participant an opportunity to experience an externally conducted research study. Thus, we conducted the semi-structured and design interviews either individually or with pairs of participants.

3.3 Analysis

The 16 hours of interview recordings were transcribed then qualitatively coded using reflexive thematic analysis [5] by two researchers. We ensured that all interviews were coded by the first

author who conducted all the interviews, and, whenever applicable, another author who observed the interview. The first author coded one transcript first, then discussed the codes with others to align on coding granularity. Each coder prioritized coding observations related to the research questions while also remaining open to capturing a broader range of potential findings. We resolved disagreements between coders through discussion.

After coding the transcripts, we conducted a bottom-up affinity diagramming process [4]. Through multiple rounds of refining and grouping, we iteratively grouped 506 unique codes into three levels of themes. The first level grouped the 506 codes into 52 themes; the next level grouped these 52 themes into 17 second-level themes; and the last level grouped these 17 themes into the three highest level themes. We organized the Findings (Sec. 4) by these three highest level themes.

3.4 Ethics and Participant Safety

We assured all participants that their participation is voluntary, all study questions were optional, their responses will remain anonymous, and they may ask us to remove content at any time after the study. To mitigate the risk that participants are identified within their agency and that agencies are identified, we do not specify participant roles beyond the participant groupings we define in Section 3.1. To mitigate the risk that public sector agencies are identified, we do not differentiate the agencies that participants are employed at and limit the amount of details we provide about each agency. This study was approved by the Institutional Review Board of the authors' home institutions.

3.5 Positionality

Our research team collectively holds expertise across human-computer interaction, computer-supported cooperative work, critical computing studies, public policy, organizational studies, machine learning, and statistics. As researchers studying public sector settings, we are conscious of our position of privilege in having research access to public sector agencies and leaders, who have opened their doors to us as external researchers. We acknowledge that our positionality and direct interactions with these agencies, along with our prior work examining the downstream impacts of public sector AI systems on social workers [7, 24, 25] and community members [47], have collectively shaped our perspectives which guide our approach to this research.

4 FINDINGS

As public sector agencies continue to develop and drop AI tools deployed in high-stakes decision-making domains, research and popular media has increasingly scrutinized their decisions [16, 41, 53]. Agencies have typically been depicted as monolithic entities making authoritative and self-assured decisions around AI. However, our findings complicate this depiction.

We unveil how agency decisions around AI design and use are shaped by agency-internal contention regarding the value and validity of AI, constant pressures from other institutions of power, and concerns around their ability to mitigate power and knowledge differentials between them and impacted communities. To contextualize agency decisions within the broader web of infrastructural and social systems they reside in, we organize our findings to center the power relations that public sector agencies hold with other sectors and stakeholders situated *within* (Section 4.1), *above* (Section 4.2), and *below* (Section 4.3) the agency.

4.1 Power Relations within the Agency: Misconstrued Concerns and Tensions around AI amongst Workers

Agency decision-makers' experiences surfaced contrasting viewpoints around the value and validity of AI among different groups of workers, including within frontline workers, within Research and Development (R&D) workers, and even within agency directors. Agency decision-makers who were proponents of increasing their agency's use of AI tools (who were the majority of the participants) felt concerned about their peers' potential misunderstandings on the role that AI would play in their agency, especially when these concerns led to the suspension of certain AI tools (Section 4.1.1). They were especially wary of frontline workers' seemingly pervasive and persistent concerns towards AI. As a result, agency decision-makers described a range of efforts to promote an organizational culture of understanding of AI, believing that disclosing how its agency would and would not use AI could help mitigate concerns. However, talking with frontline workers who were involved in sharing feedback on piloted AI tools, we observed critical gaps between frontline workers' underlying concerns and the reasons for concern that agency leaders and developers attributed (Section 4.1.2). Moreover, while agency decision-makers described that they involved frontline workers in decisions around AI design and use, frontline workers expressed that their existing pathways to inform decisions were too constrained; they desired opportunities to inform key decisions at earlier design stages (Section 4.1.3).

4.1.1 ***The value of AI tools is a point of contention among workers in agencies.*** While public sector agencies are often portrayed in the media as collective forces pushing forth the deployment of new AI tools [41], our findings suggest this depiction overlooks critical internal tensions and disagreements between agency stakeholders – including those with the power and responsibility to shape the design of these tools.

Recalling past experiences where an individual's opposition towards AI tools led to deployment halts, several Agency Leaders and R&D Workers described that one of their biggest challenges to moving forward with the use of AI-based tools was receiving "buy-in" from other leaders in their agency. This was particularly important to progressing innovations on AI in the agency, because, historically, different leaders across time and within the same agency held divergent stances around the value and validity of AI tools. Because of the positions of power that these leaders held, their stances towards AI had dramatically impacted decisions around whether to deploy AI tools within their agency. For example, one participant recalled an experience where their agency had developed a new AI tool to aid frontline workers' decisions. This tool was never deployed due to opposition from a agency director:

> "[The director] wondered what the algorithm would do if it were deployed at that point for [their family, when they were called into child services] and [their] assumption was that would not help [them] out, because [the director is] a minority, a teen [parent]. [In] reality, the algorithm probably would have really benefited her, because that would be considered an extremely low risk case." (L04)

Many other leaders, like this participant, believed that opposition questioning the fundamental value of predictive AI tools was not well-founded. The overall sentiment that there are concerns to applying AI tools to certain contexts which are under-acknowledged by most agency decision-makers resonated especially with one R&D Worker:

> "I am much more conservative about the use of algorithms, especially with communities that tend to get the short end of the stick in terms of power and decision making [...]

> I think there's lots that can be done, but I'm not sure that decision making in these contexts is how I would have used algorithms. But that's just me." (RD02)

The participant had observed several instances of their colleagues developing AI tools, and through the years, had raised concerns around *who* the AI tool was designed to benefit. For example, the participant shared concerns around how the agency prioritized creating AI tools that helped achieve their own objectives (e.g., federal outcomes) rather than objectives that would benefit community members:

> "I am here to tell you [when] these conversations have happened in agencies, the outcomes tend to be the things the agencies are interested in, not the things that the people that are being served by the agency are interested in. And so I think that's a real concern. People are tired of me saying, 'This seems to be all about us what works for us."' (RD02)

While uncommon across Agency Leaders and R&D Workers, participants described that the above concerns were widely shared across frontline workers. In the next section, we discuss how participants interpreted and responded to internal concerns around the design and use of AI tools, which often came from frontline workers.

4.1.2 ***There are disconnects between frontline workers' AI-related concerns and their concerns as construed by agency workers with higher institutional power.*** Agency decision-makers believed that agency-internal contention around their use of AI were especially pervasive amongst frontline workers. In this section, we describe how agency decision-makers' interpretations of frontline workers' concerns shaped their goals for communicating with frontline workers on topics related to AI. Importantly, throughout our interviews, we observed discrepancies between the underlying causes for concern frontline workers voiced and the perceived causes attributed by agency leaders and developers.

Agency Leaders, R&D Workers, and Frontline Leaders often described efforts to "convince" frontline workers in their agency of the value AI-based approaches could bring, to help mitigate their concerns about AI-based tools. However, they often described that their efforts to inform and align frontline workers' perspectives towards AI tools seemed futile, given the persistence of frontline worker concerns. For example, one Agency Leader explained:

> "It's not priority [for frontline workers] to be using or [seeing] value in data-driven approaches like predictive analytics [...] It is a challenge [to] convince other teams that okay, by [...] using this approach you actually stand to gain [...] xyz aspects of things, you know. If your folks, if your staff gets read up sooner or your resources get [...] more efficiently used, then it is a win-win situation. There are challenges in trying to convince non-data people that there is merit to using a data driven approach." (L03)

As another R&D Worker put it: "it seems that no matter how much we voice that that concern is present and hard for breath to convey, and have people really grasp that the tools meant to help inform their decision them with another tool in their tool belt, so to speak, rather than replace them as the decision maker" (RD04).

Beyond talking with frontline workers about the potential benefits of AI tools, some participants also believed that their AI evaluation approaches can help promote a more accurate understanding of AI across their agency. In particular, by applying model accuracy and fairness measures on human decisions, one participant hoped it would help reduce frontline workers' distrust towards AI tools by conveying similarities between human and AI decision-making:

> "I think that one of the goals, at least in my mind, is to kind of change, the perception of what an algorithm is in like a lot of the kind of non-technical portion of the agency [...] they're very on board with the idea of algorithms and fairness, without realizing that the way that they're making decisions is kind of [...] an algorithm in and of itself.' (RD05)

Importantly, by talking with frontline workers who had experience piloting new AI tools, we learned that frontline workers were concerned about a broader set of challenges that come from AI use in social services, than what agency decision-makers we talked with had acknowledged. Agency leaders and R&D workers often described the lack of technical literacy amongst frontline workers, often referring to them as "non-technical" portions of the agency, impeded their ability to seeing the value in AI tools. For example, one R&D worker described:

> "The perception of what an algorithm is in a lot of the kind of non-technical portion of the agency [...] because a lot of people who don't necessarily know the ins and outs of how machine learning works kind of [learn] through [...] popular media and stuff [...] There's almost like a mystique around machine learning algorithms, like there's some amazing thing that is all knowing and all seeing, and therefore can predict all these different things. " (RD04)

Agency decision-makers also often believed that misconceptions on how their agency would use AI, rather than potential problems with the design of an AI tool, could explain why frontline workers were concerned. For instance, the R&D Worker continued to describe frontline workers' fear towards AI as a problem of not understanding how AI will be used:

> "In general one of the things we do meet often [from frontline workers] [...] is the general distrust of algorithms and fear that it's going to replace people's jobs versus help them inform the decisions rather than take over the decision for them. [...] it seems that no matter how much we voice that, that concern is present" (RD04)

However, frontline workers we talked with emphasized concerns they had about how well agency decision-makers in their organization understood the impacts of AI tools in their work processes. They desired broader acknowledgement of the downstream consequences that deploying and maintaining an AI tool can have on frontline workers' responsibilities and labor. For example, one frontline worker expressed frustration with claims they heard that an AI model only requires workers to "push a button":

> "This tool is only going to be as good as the data. [...] How is that [going to] work? [It requires] someone that can actually monitor ensuring that the data that's being entered in is done consistently [...] What's that saying? Death by a 1,000 cuts like, you know, like folks are like. 'Oh, it doesn't hurt that hard. It's not hard to push that button, or you know. Do this, or do you know, like what's so hard about it.' Well, folks that give that kind of feedback don't understand to kind of work that we do." (F02)

They described that, because each agency has different priorities for what data to input into their system, some model features may not be consistently documented in specific agencies. For state-level AI tools, frontline workers are responsible for carefully collecting and inputting data for the model. In one case, an R&D Worker from another agency recalled a researcher flagging the added labor for frontline workers that would come with deploying their new AI tool. While there were suggestions to roll out the AI tool alongside other technical updates to help alleviate frontline

workers' workload, there were infrastructural barriers to doing so. In particular, the proposed solution-developing a system that automatically populates data fields with needed information about a given case–could not actually be implemented by the development team, because they did not have maintenance control over the data inputting system. In the next section, we elaborate on how frontline workers also expressed a desire to inform key decisions on the design and use of AI tools that would directly impact their work.

4.1.3 ***Frontline workers desire more collaborative and early-stage participation on AI design.*** Participants across agencies described various touch points with frontline leaders and workers. In a couple of agencies, Frontline Leaders were asked to check whether the most heavily weighted variables in a predictive model "makes sense" to assess its face validity. Agency decision-makers also described involving some Frontline Workers (those who are asked to use the AI tools they develop) in later stages of their development process, in particular, to pilot their tool. When we talked with the frontline workers who were involved in the piloting process, however, we learned there were challenges to having frontline workers share feedback on the AI tool. Aligning with the overall sentiment that there are major communication disjunctions between the frontline and higher ups who have more decision-making power over AI, one frontline worker shared that their peers were going through the motions of running the AI tool, but not actually critically interpreting or evaluating it. The worker hypothesized several reasons for this ranging from potential fear of insulting the AI developers, to not having enough time to notice and critically evaluate the model, to confusion around what feedback the AI developers were asking for.

Frontline Leaders and Frontline Workers we talked with desired opportunities to more proactively engage with developers on discussions surrounding the design, evaluation, and use of AI tools. In current development processes, they felt that they did not have an opportunity to inform consequential decisions around the design of the model, such as what target the model should be trained to predict, as well as organizational polices governing model use, such as whether workers can exercise discretion in the use of the model. While frontline leadership were more likely to be consulted in middle (rather than later) stages of the AI development process, one Frontline Worker described that those in leadership positions may forget what the day-to-day work looks like on-the-ground (e.g., constantly "putting out fires"), thus limiting their ability to anticipate downstream consequences from deploying models.

4.2 Power Relations with those Above: Constant Pressures and (Lack of) Support from Other Institutions of High Power

Agency leaders and developers frequently described how powerful entities outside of the agency—e.g., federal government, state courts, and contracted private companies—shape their decisions around AI tools. Whereas prior empirical research has largely focused on how agency decisions and policies have downstream impacts on communities and frontline workers, our findings surface how legal and infrastructural barriers stemming from other institutions of power impact these decisions. For example, participants expressed frustration on their ability to interrogate ethical considerations on AI adoption decisions–a practice that had been largely ignored by contracting companies who hold a dominant negotiation position and are disincentivized from addressing the agency's concerns.

4.2.1 ***Agency decision-makers that procure AI technologies felt their relations with private companies fundamentally constrained their power on ensuring responsible AI design.*** Participants working in a city-level government agency that procured AI-based tools from private companies described a range of tensions around their interactions with these companies. In these settings, participants expressed frustration that private companies that develop their procured

technologies often have minimal incentive to follow responsible and collaborative AI development processes.

Participants whose agency procured AI technologies from private companies described that they perceived power imbalances in their relationships with private companies. Participants described their relationships with private companies were often heavily shaped by the agreements listed in their procurement contracts with the company. Procurement contracts help define the conditions of the relationship between the city government department and a given private company that has agreed to develop a new technology. Participants described it can be difficult to anticipate what to include in the procurement contract before the AI tool is actually developed, leading to situations in which they have limited leverage in requesting certain model information and performance requirements. For example, one participant described an experience learning about the importance of carefully wording procurement contracts, after being denied of important information from the private company. In particular, the participant shared an anecdote about how someone in the department noticed an AI-based risk assessment tool that evaluates road conditions was disproportionately flagging roads located in affluent neighborhoods. The department requested more information on this concern from the private company, but was denied any details:

> "One of our city Council members very quickly realized that the scores were prioritizing investments in areas that already are known to be some of our wealthiest neighborhoods. [...] Once we went back to the vendor to request [an explanation], the vendor said: 'We cannot share that information with you because it's proprietary."'(L05)

Agency stakeholders believed that their ability to use stringent procurement contracts was very restricted in practice. In particular, they expressed concerns that if the language in a procurement contract is too strong, private companies could simply decide not to sign the contract and instead go to another customer where the procurement contract is more favorable to them. The participant expressed their frustration around the lack of effective incentives for private companies, describing:

> "[T]here are fundamentally no incentives, rules, or regulations to incentivize the private sector to unlock or show how those systems work. [...] I have had companies tell me directly, point blank that it is not profitable to be ethical. And because of that, they really don't have anyone holding them to the fire to actually change their business practices [...] to reduce bias or achieve some of these goals around equity." (L05)

Participants at this agency also described new challenges surfacing from technologies that were procured from private companies before AI-based technologies were widespread. Because older procured technologies followed the SaaS (Software as a Service [50]) model, the participants described that private companies often integrate AI-empowered features in their software updates without informing the technology users or city government. Without an awareness of which government technologies now use AI-powered features, participants described that workers may not be well-prepared to identify and account for possible technical failures from these updates. Therefore, one participant described that one of their projects includes documenting where AI is used across their departments.

Participants also described an overall fear of over-relying on procured technologies from the private sector in the longer-term, given that this may lead to allowing private companies to have irreplaceable power over city departments. They described that it risks allowing the vendor to have "a lock with the city," making the agency more susceptible to losing an understanding of the behind-the-scenes mechanisms around how their technologies are serving the public community.

While they desired support for crafting more power-balancing procurement contracts, participants also acknowledged that any interventions at the procurement-level could only provide a limited amount of leverage to the city government (as described above). As a result, participants desired ways to incentivize private companies to prioritize ethical considerations in the design of their algorithms. For example, one participant described that having private company vendors be fully or partially liable to lawsuits resulting from the use of AI tools could encourage them to prioritize responsible development values. Participants also desired mechanisms to share experiences and information across public sector agencies, like enforcing a certification or ratings system where agencies could signal the quality of private company vendors they worked with in the past. To evaluate the effectiveness and impacts of new AI tools in a more collaborative and iterative approach, one participant also expressed an interest in exploring the use of digital test beds.

4.2.2 ***Agency decision-makers desired federal support that is better contextualized to their actual needs and challenges.*** Participants working in a city-level government agency also described that the federal government provides support by publishing guidelines that inform their efforts towards algorithm design and evaluation. For example, they described that the NIST AI Risk and Management Framework shaped their pre-deployment evaluation priorities for AI tools:

> "The Federal Government is behind cities. So cities are really at the front lines of innovating a lot of these tools, policies, and procedures, and the Federal Government, you know, often invite cities to consult on how they should develop national strategies or frameworks for this work [...] generally they push out information, guides, etc., that cities then adopt [...] because of that Federal backing of those [NIST cybersecurity] standards, the private sector has revised their approaches in order to be in alignment with NIST standards" (L05/L06)

However, participants also described that federal guidelines often overlooked real-world resource constraints in their agency and government (e.g., by requiring creation of new roles). For instance, one participant described:

> "I find a lot of AI guidelines are very aspirational and not very realistic for municipal government, because they don't address the capacity issue and because they require in many cases the creation of new types of roles and positions within the city that don't exist in many cases. So, for example, an algorithmic auditing role: someone whose purpose within local government is to process and evaluate AI tools for their abilities to meet certain standards that also don't exist yet." (L05/L06)

Participants further imagined that federal guidelines could help align incentives towards responsible AI development between the city government and private companies. One participant explained that, while city government workers themselves have limited leverage, standards that are "backed" by the federal government are more likely to nudge private companies to adjust their development approaches. However, the participant described that federal responsible AI guidelines alone may not be enough to improve private sector companies' practices. They expressed concerns that., without proper "incentives and enforcement," guidelines could be ignored by private companies: "If you're setting a guideline for, you know, private sector companies, how realistic is it given that they're really, main incentive is the bottom line, and to remain solvent as a company" (L05/L06). Therefore, the participant advocated for federal requirements to mitigate the risk of private companies simply rerouting their services to other (less ethics-prioritizing) customers.

4.2.3 ***Agencies grapple with concerns about AI outputs being misused in state courts.*** Several participants raised concerns around how risk scores generated from AI tools could be used outside of their agency, in legal cases at state courts. Participants expressed concerns that AI risk scores could be misused or misinterpreted by attorneys in the court. For example, frontline workers imagined that attorneys might use "incorrect" risk scores to their advantage:

> "... but we were predicting that attorneys are going to say, look, [the agency] has this tool [...] that says this family would probably be just fine if they return these kids home. And meanwhile we're over there saying there's no way this family is ready. They've done nothing. They're using substances currently or [...] There's violence in the home." (F03)

Some leaders and developers also described experiences receiving advice from attorneys to avoid developing a new AI-based decision-making tool, to mitigate the potential for producing additional information that might be used against them in a lawsuit. One participant described that, in existing AI-based decision-making tools, they attempt to mitigate the risk of score misinterpretation by including a text description indicating what a risk score does and does not indicate. However, participants acknowledged the limited leverage they currently have in preventing legal misuse of AI-based risk scores.

4.2.4 ***Some agency decision-makers–but not others–want to ensure their decisions about AI are better informed by legal considerations.*** To both inform the design of AI-based tools and inform legal experts of the processes and capabilities of AI-based tools, several participants described a desire to have earlier discussions about the model with legal experts. Describing the fundamentally intertwined relationship between child welfare screening and the legal system, one frontline worker emphasized the importance of considering legal perspectives early and often: "A child welfare is based on policy, right? It's based on laws. there's no way around it" (F02). Another participant further elaborated: "Policy, procedure, statute rules [...] all need to be taken into consideration [...]. You need to be having conversations with your Department of Justice and attorneys" (RD06).

Participants described that their cognizance of how legal considerations could shape the use of AI-based tools was based off of prior experiences they had interacting with the legal system due to the design of their AI tools. Participant described that these interactions often occurred in a reactive, ad-hoc manner, for example, through lawsuits against the agency: "In our case there was also a lawsuit that was filed, which changed the way we responded to the different needs of the system" (L03). This participant imagined that, by working in collaboration with legal experts who were familiar with existing statutes and laws, they could help bridge new understanding of how these considerations play out in AI development.

However, not all participants were interested in having collaborative discussions with attorneys and legal experts. One participant, who served in a leadership role, believed that attorneys were overly "conservative" towards the use of new AI-based tools. This participant, along with another developer, was concerned that policy and legal considerations compromise the potential usefulness of developed AI tools, hampering innovation in the longer-term:

> "Well, I try and stay away from the attorneys for a lot of reasons. They're always on the Conservative side, and there's no consideration for the benefit to the family. They only consider the risks and not the benefits" (L04).

4.3 Power Relations with those Below: Barriers to Hearing and Empowering Impacted Communities

Recent research has increasingly called for community participation around the design of public sector algorithms. As a result, the participatory machine learning and human-computer interaction communities have proposed a range of methods for collaborative AI design with diverse stakeholders, including those without technical expertise (e.g., [27]). However, we know little about whether and how these design approaches and tools are actually used in practice. Furthermore, developing an understanding of the real-world factors that shape existing agency-community collaboration around AI development processes is critical to ensuring that the participatory model and design approaches can actually be transferred and used in practice.

In this section, we shed light onto infrastructural and social barriers that are often overlooked when designing approaches for community participation around pubic sector AI development. Overall, we found that individual agency leaders and R&D workers expressed interest in involving community members (Section 4.3.1) but, without formal incentives to set up these collaborations, they faced downstream challenges when attempting to advance community engagement efforts in practice. These challenges further disincentivized agencies from expanding participation around their AI development pipeline. For example, participants described a lack of infrastructural support in identifying and connecting with relevant groups (Section 4.3.2). Moreover, some agency leadership and developers were cognizant of power and knowledge differentials between the agency and community members, which they believed would challenge R&D workers' abilities to have collaborative discussions with community members around AI development (Section 4.3.3).

4.3.1 ***Agency decision-makers expressed interest and value in involving community members in decisions about AI design.*** Some participants (RD02, RD04, RD07) expressed interest in expanding their agency's currently insulated approach to developing AI tools by involving community members in their development process. For example, one participant (RD02) acknowledged that leaders and developers internal to their agency likely view, interpret, and make design decisions around the AI tool differently than those "outside" of their agency, like community members. Another participant expressed frustration that their agency appeared to develop AI tools that prioritized the agency's needs over those of the community, e.g. developing AI tools to meet federal targets that may not be aligned with community needs.

This participant further elaborated that, if public sector agencies do not begin listening to community concerns and collaborating with communities around the tools they build, then they are likely to miss opportunities to address fundamental, root problems and risk focusing a disproportionate amount of attention towards fine-tuning algorithm designs. The participant advocated for the importance of engaging community members, arguing that only by meaningfully engaging with those "closest to the problem" can agency researchers move past their own privilege and blindspots. Without engagement with communities to understand their true problems, researchers "can't get out of our own way" (RD02).

4.3.2 ***Agencies are still building out infrastructural support to sustain community partnerships around AI.*** While participants were interested in improving their agency's practices towards involving community members in AI development, they felt there was inadequate infrastructural support to do this in practice. Other agency leaders and developers anticipated potential power and cost imbalances between the agency and community, and desired ways to mitigate potential harms arising from such concerns.

Participants at different organizations had varying levels of structure in their practices for involving community members in their AI development process. At the agencies where algorithms

were developed in-house, one agency described receiving feedback on ethical considerations around developed AI tools from an external research committee, which involved parent advocates and people who have lived experience in foster care.

On the other hand, participants from another agency that develops in-house AI tools described facing several barriers to involving community members. While participants expressed interest in receiving feedback on their AI tools from community members, they described that their agency does not currently have the infrastructural systems in place to support and sustain community engagement around AI development. To learn how to establish pathways for involving community members, some participants described plans to talk with other departments in their agency (e.g., an office of equity). Participants (RD06, RD07) described that the agencies need guidance on identifying community members to reach out to, ensuring that participation is accessible, determining logistics around payment and involvement time, and establishing partnerships with mutual trust: "It's just a different piece of the puzzle here to get that input" (RD06). The participant further elaborated:

> "'It's building that network of people who want to come in and give feedback and participate in these [activities]... If we really want genuine input and people to be coming in from external to our agency, we need to make these things more accessible and more available to them. They're not during like eight to five business hours Monday to Friday. We [should] do groups after five on weekends. Just can we pay people for their involvement? These are conversations that we're having right now. So that it's not some false inclusion and access conversation. It's like true participation." (RD06)

4.3.3 ***Agency decision-makers felt ill-equipped to effectively communicate with impacted communities due to power and knowledge differentials.*** Beyond having more infrastructural support and guidance on building sustainable partnerships with community members, some participants also anticipated communication challenges with effectively integrating perspectives from those beyond their usual research and leadership team. For example, one participant elaborated that well-intended developers may struggle to understand community concerns due to lack of lived experience, while community members may not have the technical knowledge needed to inform algorithm design. This participant recalled they have often seen "perfectly well-intentioned analysts not hearing or understanding [local community members'] concerns," even when agencies make an effort to engage with local communities (RD02). The participant further worried that some of their peers may not take feedback from "non-technical" people seriously. They described that they have always expressed concerns about the agency's siloed approach to designing algorithms but felt that the agency as a whole did not prioritize addressing these concerns ("[colleagues are] tired of me saying [this]"). Another participant raised concerns about the feasibility of involving people without domain knowledge on AI in ways that bring value to the conversation:

> "I've seen a lot of the time that people who have no technical background, they may sit in this meeting right and just not have any idea of what the conversation is. When bringing everybody together, how do we communicate that both people with technical backgrounds and those with no technical background are able to follow the discussion on the conversation? I think that's hard. I think it's almost like the power differential to and from the technical folks." (L04)

A couple participants additionally acknowledged the power imbalances that might hinder effective collaboration between AI developers and community members. One participant anticipated that community members would have to endure higher emotional burdens from participating in AI design compared to public sector agency workers, as they would be "asking a lot more of

the communities we serve to come and dwell on those spaces than is asking of us" (RD07). They continued to advocate that government agency employees should take the initiatives to share the power and spaces, in order to remedy "perceived harms from governments on the people." (RD07). However, participants (including RD07) described that their agency doesn't currently have the cultural norms and guidance needed to support effective power sharing that would be required in collaborative design processes that go beyond consultation:

> "In government, to really honor [power sharing] and just not be like, you know, mining people for their opinions, but to really allow community to kind of drive some of these decisions a little bit, and determine the shape of predictive models and whatnot that sit on decision points that are important to their lives. That still feels a little foreign, and we–as in the government agency–I don't know that we have great mechanisms yet for some of that kind of power sharing." (RD07)

5 DISCUSSION

> "If we don't have accountability for the impact of those algorithms on people's lives, we get overly caught up in the ability of rules to solve problems [....] As researchers, we almost can't get out of our own way. The algorithm isn't the problem." (RD02)

5.1 Interpretative Overview of Findings

As public sector agencies rapidly create and use AI systems in high-stakes decision-making domains like social services, it becomes increasingly critical to ensure that decisions around AI design and use are well-informed, by reflecting on *who* or *what* inform these decisions, and *why*. To improve the design of AI systems, the SIGCHI research community has proposed a range of approaches to help expand stakeholder participation around the AI development pipeline. It is, of course, expected that bringing HCI methods and tools from the research community into real-world practice is a slow and challenging process. However, our findings suggest that there are unique barriers in the public sector that pose further challenges and disincentives to successful real-world adoption. Agency leaders and technology builders in our study lacked the conversational skills and resources needed to have the reflexive discussions necessary for understanding workers' concerns around AI (Section 4.1.2). Without adequate communication mechanisms in place, participants found it challenging to ensure that collaborative design approaches (e.g., participatory design)–even if adopted into practice–could be effectively used by agency workers. In public sector social services, frontline workers who are asked to use AI in their decision processes may operate in separate, siloed offices from higher-up agency leaders and technology developers who typically drive decisions related to AI design and use. Therefore, improved communication channels may be a prerequisite to introducing more collaborative AI development processes that get adopted and used effectively in practice. Importantly, even when agency decision-makers were aware of power imbalances and knowledge differentials (e.g., with impacted communities), this awareness further disincentivized agencies from integrating community participation across their AI development pipeline. Our paper contributes to the growing call from researchers to create more practical avenues to support responsible AI development practices (e.g., including examples of how to involve "diverse stakeholders" rather than assuming this is something the practitioner may already know how to do [54]).

Moreover, participants expressed frustration with the lack of realistic guidance, support, and communication from other institutions of high power, including the federal government, legal systems, and contracted private companies. Our findings point to the need for regulation and policy that could support public sector agencies in keeping contracted private companies accountable for ethical and fairness considerations in their AI development process. In the absence of such measures,

an increasing number of agencies are turning to developing AI systems in-house, to overcome incentive misalignment with contracted private companies and ensure greater transparency and autonomy over design decisions. As public sector agencies continue along this path, it is also critical to acknowledge that agencies lack the level of infrastructure (that large private companies have) to adequately support AI development, deployment, and maintenance. Overall, while prior research has tended to focus on advancing technical and design solutions to support more value-sensitive AI development practices, our findings point to a broader set of real-world considerations-–including regulatory, infrastructural, and cultural factors-–that play a crucial but largely overlooked role in many existing research contributions aiming to improve public sector AI development.

5.2 A Power-Conscious Agenda for Public Sector AI: Recommendations to Improve Agency Decisions

Based on the findings, we propose a power-conscious agenda for researchers and policymakers to better support public sector agency decisions around AI design and use. By "power-conscious," we refer to the practice of centering networks of power relations–including those that exist within agencies, between agencies and other institutions of power, and between agencies and the communities they serve–to identify how they could impact agency decisions around AI, and what concrete avenues for improvement exist. We discuss three avenues for future work, crosscutting regulatory, organizational, social, and design considerations. We elaborate on key findings and prior literature to support our recommendations.

5.2.1 ***Support reflexive conversations amongst stakeholders to understand their underlying concerns around AI.*** Our findings surface a need for more conversational tools to help surface differences in perspectives, assumptions, and experiences across different stakeholders; and bridge understanding of how these cross-stakeholder differences may shape perceptions around a given AI systems' value and validity. Without such practices in place, public sector agencies may have impoverished interpretations of workers' concerns around AI. In our study, many agency leaders and technology developers believed that frontline workers had persistent and pervasive concerns around AI systems. They speculated that frontline workers' perceptions may be heavily biased by negative media attention around AI systems or their lack of technical expertise to understand the value of AI systems (Section 4.1.2). While media exposure and technical expertise may play a role, our findings also suggest that there is a broader communication gap underlying the relationship between frontline workers and agency workers with greater institutional power (e.g., agency leaders and technology developers). Frontline workers wished for a broader acknowledgement of the downstream consequences of AI deployments from those creating AI tools, but they felt that those who had a voice in the development process either had no experience doing their work or had spent too much time away from it to remember their day-to-day needs and challenges.

Recent research has described a growing trend of AI deployments in social services overlooking the underlying needs of frontline workers. For example, prior work has also observed that frontline workers benefit from better understanding their own decision-making tasks and goals, allowing them to identify misalignments between how they were trained to make decisions and what forms of support an AI tool provides [24, 55]. It is worth noting that in social services–much like in other domains including education and home healthcare–frontline workers have historically been devalued and underpaid, and, to this day, the workforce consists largely of women and people of color [39]. Collectively, these factors pose additional and unique challenges to ensuring workers across an agency's organizational ladder have their perspectives and concerns adequately heard and addressed through reflexive conversations.

To improve the design of AI tools in these domains, workers within organizations need support in understanding one another's perspectives and concerns–even before attempting to expand participation in their AI development process. Researchers can help innovate new approaches and methods to scaffold conversations that more effectively surface and address the underlying concerns around AI across stakeholders with disparate educational backgrounds, institutional power, and experiences. For instance, while existing approaches and methods in research typically assume the presence of a facilitator, it is not entirely clear whether agency workers–who are trained in other specific skills (e.g., social services, management, technology development) but not in reflexivity–are best positioned to facilitate such nuanced conversations. There is an opportunity for future research to help address this gap, alongside relevant regulatory and policy efforts as described in the next section.

5.2.2 ***Provide more context-aware federal and regulatory support, including the creation of new roles for responsible AI.*** Our analysis surface a need for more federal support to help public sector agencies allocate new roles and resources dedicated to supporting responsible AI development practices. Participants across the agencies emphasized that they were constantly busy ("putting out fires") and under-resourced, limiting their ability to allocate adequate time or resources toward responsible development efforts (Section 4.1.3). Some participants described how existing support from the federal government (e.g., NIST AI RMF [48]) was overly idealistic, for example, by requiring new roles to be created in order to follow them. Adopting "best practices" proposed by the federal government or civil society (e.g., AI Impact Assessments [31, 38]) may require agencies to shift their allocation of already-strained resources, further disincentivizing agencies from adopting these proposals.

Relatedly, participants expressed uncertainty and lack of guidance around how to do the work of community outreach and engagement, especially given knowledge of power differentials and "years of perceived harm from the government to the people" (Section 4.3.3). The challenging task of navigating these nuanced social situations further disincentivized agencies' progress on expanding community participation in their AI development process. While having practical guides and tools to scaffold more participatory practices (as elaborate in Recommendation 5.2.1 and 5.2.3) may help, it may additionally be helpful to have specified roles and experts dedicated to supporting these responsible development practices. For example, large technology companies with more resources have turned to creating new roles who are responsible for helping product teams effectively use artifacts and processes designed to help mitigate bias or improve transparency in development (e.g., Responsible AI Champs at Microsoft [34]). These roles are intended to help product teams *learn* how to do responsible AI work, by bringing in individuals with relevant training and expertise. Our findings surface how, without having such roles dedicated to supporting responsible development practices, individual workers who were personally motivated to address ethical concerns around AI currently bear the burden of advocating for and educating their colleagues on related topics (reflecting similar trends as those documented in the private sector [21, 29]). Participants described how they urged their colleagues to reflect on critical questions like who their technologies were intended to benefit (if not the community); however, they felt their perspectives and concerns were overlooked or dismissed by their colleagues. Institutionalizing this (currently invisible) labor through the creation of new responsible technology development-related job roles could help legitimize these efforts, supporting both these voluntary advocates and the agency's overall efforts towards practicing responsible development. Again, we emphasize, though, that creating such roles will require additional resources and regulatory support.

Our findings surface a need for regulation that supports public sector agencies in interrogating ethical considerations in the technologies they adopt from contracted companies. One example of

such a regulation might require private companies to disclose their AI design and development method, evaluation approach, and specific evaluation outcomes. In some cases, public sector agencies are themselves limited in their ability to make value-aligned decisions. Procurement contracts don't support the iterative nature of harms identification and mitigation in responsible AI pipelines, so agencies hoping to mitigate potential harms from their AI systems may have limited capacity to contribute to such efforts. In our study, participants struggled to preempt every possible harm, because responsible AI necessarily involves iteratively learning and evaluating across the AI development process. While the academic community has established how private companies in the U.S. struggle to establish a culture to prioritize responsible development practices, there has been inadequate attention paid to how these challenges may ripple down to impact the public sector's decisions and practices around AI. Future research should continue to understand the relationships between public sector agencies and contracted private companies creating AI systems.

5.2.3 ***Create more practical approaches for non-technical stakeholder participation around AI***. Our findings point to a critical need for more *easily adoptable* approaches, methods, and tools that expand "non-technical" stakeholder participation across all stages of the AI development process. Participants acknowledged that their own R&D teams' practices were insulated from the broader community, that their experiences were not representative of those they serve, and that they desired ways to expand participation around their AI development process. These sentiments all point to a desire for improved, meaningful community engagement. Yet, participants described minimal progress in their efforts to do so, and were skeptical of the feasibility of such approaches–at times, mistakenly or without evidence. For example, some participants believed that expanding community participation in development processes may be less valuable in early-stage design. However, prior work has shown that many of the more consequential insights offered by community members target ideation in early problem formulation stages. Prior research has demonstrated how individuals without a background in AI (including frontline workers [25] and impacted community members [8, 27]) have complementary knowledge to technology workers, enabling them to help improve the design of AI systems.

Importantly, recent research has surfaced the lack of concrete guidance on how to complete the *social* tasks necessary in practices like participatory AI design; most responsible development-related methods and approaches available today instead opt to provide guidance on completing technical tasks [54]. Beyond ensuring practical feasbibility (e.g., by providing concrete guidance on how to connect and form a working relationship with a given stakeholder), it would be helpful to have more publicly available examples demonstrating the value of participatory approaches. Future work documenting end-to-end case studies of participatory design approaches implemented in the public sector could help demonstrate the value and practicality of participatory approaches, and distill lessons for future adoptions of such practices. As described in Recommendation 5.2.2, successful adoption of these approaches may necessitate the creation of new roles trained to help bridge responsible practice into the public sector.

5.3 Limitations and Reflections

In this paper, our goal was to understand how U.S. public sector agencies, particularly those providing social services, make decisions around the design and use of AI systems. To meet this goal, we interviewed participants working in public sector social service departments in the United States. Given differences in existing regulatory requirements, cultural norms, and other country-specific factors that impact AI design and development practices, our findings are more likely to generalize to similar agencies within the United States, not necessarily to those in other countries (as evident by findings from [52]). We also acknowledge that public sector agencies within the

United States may differ in size, access to resources, compliance requirements, leadership inclination towards investing in AI systems, and many additional factors. Based on the authors' informal conversations with other public sector agency leaders across the United States and prior literature, we believe that the agencies we studied may be relatively advanced (i.e., ahead of the curve) in thinking about AI applications and responsible development practices. Therefore, it is possible that the challenges discussed in this paper may arise even in public sector agencies where there is considerable investment on supporting responsible AI practices. Moreover, given the nature of challenges described in the paper (focusing on infrastructural, legal, and social challenges), and the generality of their sources across public sector social service departments (e.g., being under-resourced, having historically tense relations with the community), we believe it is possible that other U.S. public sector social service agencies may encounter similar barriers and challenges as those described in the paper.

The act of empirically *studying up* public sector AI involves understanding barriers and challenges from the perspectives of those who hold significant power and responsibility over their existence. To help scaffold relevant reflections and deliberations around this potentially sensitive topic, our study included a design activity that asked participants to think about a hypothetical scenario considering the use of AI and asked them to describe their past experiences to help reason about their responses. However, it is possible that participants were not comfortable being fully transparent about their experiences and challenges. Moreover, participants may be self-selected to disproportionately include individuals who are interested in improving their or their agency's practices around AI. A critical line of future work is triangulating observations from this and other research studies with other relevant stakeholders, including frontline workers, impacted communities, and legal experts.

Finally, agency decision-making is often made "public" through online documentation (e.g., [17]) but a lot is still excluded from those documents. Understanding the latent perceptions, assumptions, and experiences underlying agency decision-making is crucial to improving development practices in the public sector. We are deeply thankful to the participants and their agencies for sharing their time and experiences with us. We hope that collaborations between public sector agencies and independent researchers (such as those that made this study possible) become more commonplace.

ACKNOWLEDGMENTS

Once again, thank you to the study participants for generously sharing their time, experiences, and thoughtful reflections with us. Thank you to our paper reviewers for their insightful feedback that helped improved this work. This work (as presented to EAAMO '23) is a work-in-progress, and we would truly appreciate any feedback others may have around how this work can be improved.

REFERENCES

[1] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 2–11.

[2] Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. 2020. Studying up: reorienting the study of algorithmic fairness around issues of power. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 167–176.

[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

[4] Hugh Beyer and Karen Holtzblatt. 1999. Contextual design. *interactions* 6, 1 (1999), 32–42.

[5] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.

[6] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. In *Proceedings of the 2019 CHI Conference on Human Factors in*

Computing Systems. 1–12.

[7] Hao-Fei Cheng, Logan Stapleton, Anna Kawakami, Venkatesh Sivaraman, Yanghuidi Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions. In *CHI Conference on Human Factors in Computing Systems*. 1–22.

[8] Hao-Fei Cheng, Logan Stapleton, Ruiqi Wang, Paige Bullock, Alexandra Chouldechova, Zhiwei Steven Steven Wu, and Haiyi Zhu. 2021. Soliciting stakeholders' fairness notions in child maltreatment predictive systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.

[9] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*. PMLR, 134–148.

[10] Amanda Coston, Anna Kawakami, Haiyi Zhu, Ken Holstein, and Hoda Heidari. 2022. A Validity Perspective on Evaluating the Justified Use of Data-driven Decision-making Algorithms. *arXiv preprint arXiv:2206.14983* (2022).

[11] Kate Crawford. 2021. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

[12] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2021. Stakeholder Participation in AI: Beyond" Add Diverse Stakeholders and Stir". *arXiv preprint arXiv:2111.01122* (2021).

[13] Andy Dow, Rob Comber, and John Vines. 2018. Between grassroots and the hierarchy: Lessons learned from the design of a public services directory. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

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

[15] Diana Forsythe. 2001. *Studying those who study us: An anthropologist in the world of artificial intelligence*. Stanford University Press.

[16] Marissa Gerchick, Tobi Jegede, Tarak Shah, Ana Gutierrez, Sophie Beiers, Noam Shemtov, Kath Xu, Anjana Samant, and Aaron Horowitz. 2023. 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*. 1292–1310.

[17] Jeremy D Goldhaber-Fiebert and Lea Prince. 2019. Impact evaluation of a predictive risk modeling tool for Allegheny county's child welfare office. *Pittsburgh: Allegheny County* (2019).

[18] Kenneth Holstein and Shayan Doroudi. 2019. Fairness and equity in learning analytics systems (FairLAK). In *Companion proceedings of the ninth international learning analytics & knowledge conference (LAK 2019)*. 1–2.

[19] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. 2017. Intelligent tutors as teachers' aides: exploring teacher needs for real-time analytics in blended classrooms. In *Proceedings of the seventh international learning analytics & knowledge conference*. 257–266.

[20] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. 2018. Student learning benefits of a mixed-reality teacher awareness tool in AI-enhanced classrooms. In *International conference on artificial intelligence in education*. Springer, 154–168.

[21] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–16.

[22] Naja Holten Møller, Irina Shklovski, and Thomas T. Hildebrandt. 2020. Shifting concepts of value: Designing algorithmic decision-support systems for public services. *NordiCHI* (2020), 1–12. https://doi.org/10.1145/3419249.3420149

[23] Esther Y Kang and Sarah E Fox. 2022. Stories from the Frontline: Recuperating Essential Worker Accounts of AI Integration. In *Designing Interactive Systems Conference*. 58–70.

[24] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support. In *CHI Conference on Human Factors in Computing Systems*. 1–18.

[25] Anna Kawakami, Venkatesh Sivaraman, Logan Stapleton, Hao-Fei Cheng, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. "Why Do I Care What's Similar?" Probing Challenges in AI-Assisted Child Welfare Decision-Making through Worker-AI Interface Design Concepts. In *Designing Interactive Systems Conference*. 454–470.

[26] Rob Kitchin and Tracey Lauriault. 2014. Towards critical data studies: Charting and unpacking data assemblages and their work. (2014).

[27] Tzu-Sheng Kuo, Hong Shen, Jisoo Geum, Nev Jones, Jason I Hong, Haiyi Zhu, and Kenneth Holstein. 2023. Understanding Frontline Workers' and Unhoused Individuals' Perspectives on AI Used in Homeless Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[28] Karen Levy, Kyla E Chasalow, and Sarah Riley. 2021. Algorithms and Decision-Making in the Public Sector. *Annual Review of Law and Social Science* 17 (2021), 1–38.

[29] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[30] Milagros Miceli, Julian Posada, and Tianling Yang. 2022. Studying up machine learning data: Why talk about bias when we mean power? *Proceedings of the ACM on Human-Computer Interaction* 6, GROUP (2022), 1–14.

[31] Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. 2021. Assembling accountability: algorithmic impact assessment for the public interest. *Available at SSRN 3877437* (2021).

[32] Laura Nader. 1972. Up the anthropologist: Perspectives gained from studying up. (1972).

[33] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[34] Tim O'Brien, Steve Sweetman, Natasha Crampton, and Venky Veeraraghavan. 2020. A Model for Ethical Artificial Intelligence. In *World Economic Forum*, Vol. 14. 2020.

[35] Samir Passi and Solon Barocas. 2019. Problem formulation and fairness. In *Proceedings of the conference on fairness, accountability, and transparency*. 39–48.

[36] Sofie Pilemalm. 2018. Participatory design in emerging civic engagement initiatives in the new public sector: Applying PD concepts in resource-scarce organizations. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 1 (2018), 1–26.

[37] Inioluwa Deborah Raji, I Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The fallacy of AI functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 959–972.

[38] Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker. 2018. Algorithmic Impact Assessments: A Practical Framework for Public Agency. *AI Now* (2018).

[39] Hye Jin Rho, Hayley Brown, and Shawn Fremstad. 2020. A basic demographic profile of workers in frontline industries. *Center for economic and policy research* 7, 10 (2020).

[40] Samantha Robertson, Tonya Nguyen, and Niloufar Salehi. 2021. Modeling assumptions clash with the real world: Transparency, equity, and community challenges for student assignment algorithms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[41] Anjana Samant, Aaron Horowitz, Kath Xu, and Sophie Beiers. 2021. Family surveillance by algorithm: The rapidly spreading tools few have heard of. American Civil Liberties Union (ACLU)(2021).

[42] Devansh Saxena, Karla Badillo-Urquiola, Pamela J Wisniewski, and Shion Guha. 2020. A human-centered review of algorithms used within the US child welfare system. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[43] Devansh Saxena, Karla Badillo-Urquiola, Pamela J Wisniewski, and Shion Guha. 2021. A framework of high-stakes algorithmic decision-making for the public sector developed through a case study of child-welfare. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–41.

[44] Nick Seaver. 2014. Studying up: The ethnography of technologists. *Ethnography Matters* 10 (2014).

[45] Aaron Shapiro. 2017. Reform predictive policing. *Nature* 541, 7638 (2017), 458–460.

[46] C. Estelle Smith, Bowen Yu, Anjali Srivastava, Aaron Halfaker, Loren Terveen, and Haiyi Zhu. 2020. Keeping Community in the Loop: Understanding Wikipedia Stakeholder Values for Machine Learning-Based Systems. *Conference on Human Factors in Computing Systems - Proceedings* (2020), 1–14. https://doi.org/10.1145/3313831.3376783

[47] Logan Stapleton, Min Hun Lee, Diana Qing, Marya Wright, Alexandra Chouldechova, Ken Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. Imagining new futures beyond predictive systems in child welfare: A qualitative study with impacted stakeholders. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1162–1177.

[48] Elham Tabassi. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). (2023).

[49] Sarah Tan, Julius Adebayo, Kori Inkpen, and Ece Kamar. 2018. Investigating human+ machine complementarity for recidivism predictions. *arXiv preprint arXiv:1808.09123* (2018).

[50] Mark Turner, David Budgen, and Pearl Brereton. 2003. Turning software into a service. *Computer* 36, 10 (2003), 38–44.

[51] Elmira van den Broek, Anastasia Sergeeva, and Marleen Huysman. 2020. Hiring algorithms: An ethnography of fairness in practice. *40th International Conference on Information Systems, ICIS 2019* (2020).

[52] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–14.

[53] Angelina Wang, Sayash Kapoor, Solon Barocas, and Arvind Narayanan. 2022. Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy. *Available at SSRN* (2022).

[54] Richmond Y Wong, Michael A Madaio, and Nick Merrill. 2023. Seeing like a toolkit: How toolkits envision the work of AI ethics. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–27.

[55] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting intelligent decision support into critical, clinical decision-making processes. *Conference on Human Factors in Computing Systems - Proceedings* (2019).

https://doi.org/10.1145/3290605.3300468 arXiv:1904.09612

[56] Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. 2018. Artificial intelligence in healthcare. *Nature biomedical engineering* 2, 10 (2018), 719–731.

[57] Angie Zhang, Olympia Walker, Kaci Nguyen, Jiajun Dai, Anqing Chen, and Min Kyung Lee. 2023. Deliberating with AI: Improving Decision-Making for the Future through Participatory AI Design and Stakeholder Deliberation. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–32.

[58] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–23.

[59] Alexandra Zytek, Dongyu Liu, Rhema Vaithianathan, and Kalyan Veeramachaneni. 2021. Sibyl: Understanding and addressing the usability challenges of machine learning in high-stakes decision making. *IEEE Transactions on Visualization and Computer Graphics* (2021).