
Designing Pro-Worker Futures for AI in Socially Complex Work

Author:

Anna KAWAKAMI

CMU-HCII-26-115

August 2026

Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee:

Kenneth Holstein (co-chair)	Carnegie Mellon University
Haiyi Zhu (co-chair)	Carnegie Mellon University
Jodi Forlizzi	Carnegie Mellon University
Alexandra Chouldechova	Abridge
Krzysztof Gajos	Harvard University

*Submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy
in human-computer interaction*

This research was sponsored by the NSF GRFP, NSF FAI, K&L Gates Presidential Fellowship, NIST, Google Academic Research Award, Toyota Research Institute, and PwC.

keywords: human-AI interaction, responsible AI, future of work, participatory research and design, AI-augmented work, AI measurement and governance

Abstract

AI systems are increasingly used to assist socially complex forms of human work, centered on reasoning about and responding to the needs, behaviors, and intentions of other people. For example, AI-based tools are designed to support care workers in providing socio-emotional support or to support social service workers in making decisions about how best to allocate scarce resources. However, in practice, such systems often fail to meaningfully augment worker abilities. How can we design towards more *pro-worker futures for AI* in socially complex work, where AI deployments truly complement and enhance worker expertise and capabilities? In my research, I explore this question in close collaboration with organizations and workers in *feminized labor* contexts such as social work and education, where worker expertise has historically been devalued and misunderstood.

In the first thread of my research, I conduct field observations, design research, experimental studies, and case study analyses to advance our understanding of *why* AI for socially complex work often fails to meaningfully support workers. Across a series of studies, I adopt a social ecological approach to explore this question at multiple levels, investigating worker-level, organization-level, and societal-level challenges, and I identify opportunities for future work to address these challenges. In the second thread of my research, I explore the design of new processes and tools to begin addressing these opportunities. First, I explore whether the design of new training interfaces might help social workers learn to use AI tools more critically. Second, I collaborate with public sector agencies and community groups across the US to develop a new toolkit to support multi-stakeholder deliberations around proposed AI projects. Finally, I collaborate with a local school social work organization to explore how to support worker participation in the design of AI measurement instruments (e.g., benchmarks and metrics) that capture workers' beliefs about what it means to meaningfully augment their work.

In sum, this thesis provides conceptual and empirical foundations to guide towards *pro-worker futures for AI* in the workplace—specifically by identifying challenges and opportunities across multiple levels of the social ecology that interacts with workplace AI.

Publications

This thesis is based on the following publications. While I use the first person (e.g., “I”) to describe my work in this document, the presented work was shaped by strong, cross-disciplinary research collaborations. In particular, the research described in this thesis was completed in collaboration with researchers at Carnegie Mellon’s Human-Computer Interaction Institute, Language Technologies Institute, Machine Learning Department, Heinz College Schools of Public Policy, Information Systems, and Management, as well as researchers at Google Research and Toyota Research Institute. Below, an asterisk (*) signals co-first authorship.

Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. “Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, And Desires For Algorithmic Decision Support.” In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*.

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

Anna Kawakami*, Venkatesh Sivaraman*, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. ““Why Do I Care What’s Similar?” Probing Challenges In AI-Assisted Child Welfare Decision-Making Through Worker-AI Interface Design Concepts.” In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*.

Anna Kawakami, Amanda Coston, Hoda Heidari, Kenneth Holstein, and Haiyi Zhu. “Studying Up Public Sector AI: How Networks of Power Relations Shape Agency Decisions Around AI Design and Use.” In *Proceedings of the ACM on Human-Computer Interaction* 8, no. CSCW2 (2024).

Anna Kawakami, Jordan Taylor, Sarah Fox, Haiyi Zhu, and Kenneth Holstein. “AI Failure Loops in the Workplace: The Confluence of Overconfidence in AI and Underconfidence in Worker Expertise.” In *Big Data & Society* 2026. Also in *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*.

Anna Kawakami, Luke Guerdan, Yanghuidi Cheng, Kate Glazko, Matthew Lee, Scott Carter, Nikos Arechiga, Haiyi Zhu, and Kenneth Holstein. “Training Towards Critical Use: Learning To Situate AI Predictions Relative To Human Knowledge.” In *Proceedings of the 2023 ACM Collective Intelligence Conference*.

Anna Kawakami, Amanda Coston, Haiyi Zhu, Hoda Heidari, and Kenneth Holstein. “The Situate AI Guidebook: Co-designing a Toolkit to Support Multi-stakeholder, Early-stage Deliberations Around Public Sector AI Proposals.” In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.

Anna Kawakami, Chloe Zhao, Renee Shelby, Fernando Diaz, Haiyi Zhu, and Kenneth Holstein. “‘I want to be pushed, I want to grow’: Enabling social workers to design evaluations of LLM augmentation in their work.” In *Proceedings of the 2026 AAAI/ACM Conference on AI, Ethics, and Society*.

Acknowledgements

The acknowledgements section is arguably the most challenging section to write in the thesis. It's an impossible task: Express your gratitude for all the people that have enabled you to stand where you are today. This is my attempt at an acknowledgements section—but if you really want to hear the role you played in my PhD journey, ask me directly!

First and foremost, I would like to thank my co-advisors, Ken Holstein and Haiyi Zhu, for all the time, expertise, care, and love they have extended to me in the past five years. You have had an outsized impact on me, not just as a researcher, but as a young adult navigating life. Ken, I truly do not know where I would be today without you. You have been a shining light gently guiding me throughout my PhD. Thank you for your incredibly thoughtful approach to mentorship and advising, which always prioritized my longer-term growth as a researcher and as a person. Haiyi, thank you for supporting me in pursuing my passions throughout the past five years. I am so grateful for the expertise and genuine mentorship you've extended to me through countless meetings and projects.

I'd next like to thank my wonderful thesis committee members (Alexandra Choudhova, Krzysztof Gajos, and Jodi Forlizzi) for extending their time and expertise to me. Thank you, Alex, for being one of my life role models; I will forever be in awe of how cool and sharp you are. Thank you, Krzysztof, for adopting me as one of your PhD students (according to me); I am such a big fan of you and your students, and I aspire to one day mentor a group of students as vivacious, kind, and creative as yours. And thank you, Jodi, for always posing thought-provoking, smart questions about my research to help me carefully think through what I do and why.

Thank you to the many workers that have extended their time, expertise, curiosity, and trust towards me in participating in our research. Special thank you to the many people across various state and local government agencies, and to the wonderful group of people at System 123 for welcoming me into their community.

One of the greatest gifts of doing the PhD is the community of mentors, friends, and colleagues that it has given me. Thank you to my many mentors and co-authors—Koustuv Saha, Hoda Heidari, Sarah Fox, Devansh Saxena, Fernando Diaz, Amanda Coston, Renee Shelby, Luke Guerdan, Daricia Wilkinson, Steven Wu, Shion Guha, Matt Lee, Scott Carter, Vera Liao, Joy Ming, Emily Tseng, Chloe Zhao, Alexandra Olteanu, Kimberly Truong, Nari Johnson, Erina Moon, Rachel Kim, Megan Li—that have helped imbue joy, ambition, and growth into my research journey. A very special thank you to my undergraduate advisor, Eni Mustafaraj, for introducing me to computing research and sensitizing me to the idea of doing a PhD.

Thank you to my dearest friends, Jane Hsieh, Seyun Kim, and Tzu-Sheng Kuo, for being there for me through every high and low of my PhD. Many of my fondest memories of the PhD are from everyday conversations we've had together—many of them filled with laughter, some filled with tears, and all filled with a deep sense of fulfillment from having found people that truly care about, understand, and support you. Thank you to my broader community of friends, all of whom have been a source of inspiration and have enriched my life in some meaningful way—Lisa Orii, Kimi Wenzel, Angie Zhang, Hongjin Lin, Hugo Simão, Vedant das Swain, Sachin Pendse, Shaily Bhatt, Zana Bucinca, Frederic Gmeiner, Alicia Devrio, Ji-Youn Jung, Frank Elavsky, Jianna So, Jordan Taylor, Pranav Khadpe, Jianna So, Sohini Upadhyay, Matt Jörke, Hao-Fei Cheng, Helena Vasconcelos, Anna Fang, Ezra Awumey, Howie Wang, Logan Stapleton, and more. Y'all are some of the most interesting, quirky, and lovable people I've met in my life.

Finally, thank you to my family for trusting me to carve out my own path and pursue a PhD. Thank you to Luis for probably extending my lifespan by making me laugh every day since I've met you, and for always going above and beyond to support me. And last but not least, thank you to the goodest boy ever, Albert, for the many cuddly co-writing sessions, nice walks out in the sun (and snow), and for reeling in new friends (both human and dog) through your cuteness and playfulness.

Contents

Acknowledgements	vii
1 Introduction	1
1.1 Understanding why AI systems often fail to meaningfully support workers in socially complex work	1
1.2 Designing processes and tools to engage worker expertise in AI development and deployment decisions	3
1.3 Summary	6
I Chapters 2 - 4: Understanding AI Failures	7
2 Worker-AI Interaction	9
Part 1: Understanding How Workers Make AI-Assisted Decisions	11
2.1 Introduction	12
2.2 Background and Related Work	13
2.2.1 Designing effective human-AI partnerships	13
2.2.2 Algorithmic decision support in child welfare	14
2.2.3 The Allegheny Family Screening Tool and workplace context	15
2.3 Methods	16
2.3.1 Contextual inquiries	17
2.3.2 Semi-structured interviews	17
2.3.3 Analysis	18
2.3.4 Ethics and Participant Safety	19
2.3.5 Positionality	19
2.4 Results	20
2.4.1 How workers calibrate their reliance on the AFST based on knowledge beyond the model	20
2.4.2 How workers' beliefs about the AFST shape and are shaped by their day-to-day use of the algorithm, in the absence of detailed training	22
2.4.3 Beyond "trust" in AI: Influences of organizational pressures and incentives	28
2.4.4 Navigating value misalignments between algorithmic predictions and human decisions	30
2.5 Discussion	32
2.6 Design Implications	36
2.7 Conclusion	38
Part 2: Understanding How Workers Reduce Racial Disparities in AI-Assisted Decisions	39
2.8 Introduction	40
2.9 Related Work	42
2.9.1 Algorithm-in-the-loop decision-making	42

2.9.2	Risk Assessment Tools in Child Welfare	43
2.10	Methods	44
2.10.1	Study Context	44
2.10.2	Data analysis	45
2.10.3	Contextual Inquiry and Interviews	46
2.10.4	Positionality	48
2.11	Decision-making paradigms and terminology	48
2.11.1	Decision-making paradigms	48
2.11.2	Definitions of Black and white	49
2.11.3	Evaluation Metrics	49
2.12	How AFST-assisted worker decisions affected screen-in rate disparities	50
2.12.1	Black-white disparities between AFST-only decisions and workers' final screening decisions	50
2.12.2	Qualitative: How workers achieved lower Black-white screen-in rate disparity?	53
2.13	How worker-AFST collaboration affected accuracy and accuracy disparities	58
2.13.1	AFST-only screening decisions were more accurate than workers when measured on outcomes that the AFST was trained to predict	58
2.13.2	Qualitative: Workers and the AFST do not agree on prediction outcomes and accuracy measures	60
2.14	Discussion	62
2.14.1	Interpretation of the findings	62
2.14.2	Design Implications	64
2.14.3	Limitations & Future Work	65
	Part 3: Exploring Worker-AI Interface Design Concepts	69
2.15	Introduction	70
2.16	Background and Related Work	71
2.16.1	Designing interfaces to support AI-assisted decision-making	71
2.16.2	Challenges in ADS for social decision-making	71
2.17	Study Context	72
2.17.1	The Allegheny Family Screening Tool (AFST) and Workplace Context	72
2.17.2	Prior Studies on the AFST	73
2.18	Methods	74
2.18.1	Design concept generation	75
2.18.2	Analysis	75
2.18.3	Ethics and Participation	76
2.18.4	Positionality	76
2.19	Design Concepts	76
2.19.1	Discrepancies between human and AI decision-making	76
2.19.2	Enabling worker-AI dialogue	77
2.19.3	Communicating uncertainty	79
2.19.4	Measurement and feedback on historical decisions	80
2.20	Results	82
2.20.1	Discrepancies between human and AI decision-making	82
2.20.2	Enabling worker-AI dialogue	84
2.20.3	Communicating uncertainty	86
2.20.4	Measurement and feedback on historical decisions	87
2.21	Discussion	90

2.21.1	Concept Implications	90
2.21.2	Broader Implications	94
2.21.3	Limitations	95
2.22	Conclusion	96
3	Organization-AI Interaction	97
3.1	Introduction	98
3.2	Background and Related Work	100
3.2.1	AI in Public Sector Social Services	100
3.2.2	Studying Up Algorithmic Systems	101
3.3	Methods	102
3.3.1	Public Sector Agencies and Participants	102
3.3.2	Semi-Structured Interviews and Design Activity	103
3.3.3	Analysis	103
3.3.4	Ethics and Participant Safety	104
3.3.5	Positionality	104
3.4	Findings	104
3.4.1	Power Relations within the Agency: Misconstrued Concerns and Tensions around AI amongst Workers	105
3.4.2	Power Relations with those Above: Constant Pressures and (Lack of) Support from External Institutions	109
3.4.3	Power Relations with those Below: Barriers to Hearing and Empowering Impacted Communities	113
3.5	Discussion	116
3.5.1	Power-Conscious Implications for Designing Public Sector AI	116
3.5.2	Limitations and Reflections	119
4	Society-AI Interaction	121
4.1	Introduction	121
4.2	Addressing Challenges in the Deployment of Workplace Technologies	123
4.3	Devaluation of Feminized Labor	124
4.4	Methods	124
4.4.1	Team Experience and Case Selection	124
4.4.2	Literature Review and Analysis	125
4.5	<i>AI Failure Loops</i>	126
4.6	Case Studies	127
4.6.1	AI-Based Risk Assessment Tools Have Amplified Misconcep- tions of Social Workers' Expertise	127
	Context	127
	AI Failures in Social Work	128
4.6.2	"AI for Healthcare" Has Left Behind Devalued Healthcare Ex- perts	129
	Context	129
	AI Failures in Healthcare	130
4.6.3	Persistent Visions of Automation in Education Have Further Devalued Teachers' Expertise	132
	Context	132
	AI Failures in K-12 Teaching	133
4.7	Discussion: Toward Pro-Worker AI	134
4.7.1	(How) can design processes manipulate power?	134
4.7.2	(Why) do we need to measure human performance against AI?	135

4.7.3	What happens after worker-centered design?	136
4.7.4	The Case of Feminized Labor	136
4.8	Conclusion	137
II	Chapters 5 - 7: Designing Interventions	139
5	Improving AI Use	141
5.1	Introduction	142
5.2	Background and Research Questions	144
5.2.1	Training for AI-Assisted Decision-Making	144
5.2.2	Potential Sources of Human-AI Complementarity in Decision-Making	145
5.2.3	Research Questions	146
5.3	Methods	146
5.3.1	Feedback Design	147
5.3.2	Context	147
5.3.3	Study Design and Materials	148
5.3.4	Case Selection	152
5.3.5	Participants and Recruitment	153
5.3.6	Measurement and Analysis	154
	Standard accuracy measure: Proxy-based accuracy	155
5.3.7	Positionality	155
5.4	Results	155
5.4.1	Model Agreement: Participants learn to disagree with AI predictions more often	157
5.4.2	Worker-based Accuracy: Participants become <i>more</i> aligned with experienced workers	158
5.4.3	Guessing Error: Participants learn to make improved predictions of the AI model's behavior	159
5.4.4	Process Analyses: Participants learn to make decisions more like experienced workers	160
5.4.5	Proxy-based Accuracy: Participants become <i>less</i> aligned on the model's targeted proxy	162
5.5	Discussion	163
5.5.1	Evidence of <i>critical use</i> from training	163
5.5.2	Why did explicit feedback on AI-assisted decisions not enhance learning?	164
5.5.3	On the absence of domain expertise effects	165
5.5.4	Generalizability of design decisions and findings	166
6	Improving AI Adoption and Governance	167
6.1	Introduction	168
6.2	Background	171
6.2.1	Public Sector AI and Overcoming AI Failures	171
6.2.2	Toolkits for Responsible AI Governance	172
6.3	Methods: Co-Design and Validation of the <i>Situate AI</i> Guidebook	173
6.3.1	Participants and Recruitment	173
6.3.2	Iterative Co-Design and Validation	174
6.3.3	Qualitative Analysis	177
6.4	The <i>Situate AI</i> Guidebook	178

6.4.1	Guiding Design Principles	178
6.4.2	Content Design: Scaffolding Reflexive Deliberation	179
	Goals and Intended Use	179
	Societal and Legal Considerations	181
	Data and Modeling Constraints	183
	Organizational Governance Factors	184
6.4.3	Process Design: Designing for Practicality and Adaptability	185
	Proposed Use Case: Using the Guidebook to Support Struc- tured and Iterative Deliberations	186
	Empowering Participation: Accounting for Organizational Power Dynamics.	188
6.4.4	Success Criteria	190
6.5	Discussion	192
6.5.1	Overcoming Low Adoption Rates for Responsible AI Toolkits in the Public Sector	192
6.5.2	Expanding the <i>Situate AI</i> Guidebook to Engage Community Members	194
6.5.3	Exploring how the <i>Situate AI</i> Guidebook Can Support Delib- eration in Non-Public Sector Contexts	194
6.6	Conclusion	195
7	Improving AI Measurement	197
	Part 1: Framework for AI Measurement as a Collaborative Design Practice	199
7.1	Introduction	200
7.2	Background	201
	7.2.1 AI and Measurement	201
	7.2.2 Stakeholder Participation and Measurement Design	202
7.3	Measurement as Collaborative Design	203
7.4	Methods	206
	7.4.1 Corpus Development	206
	7.4.2 Analysis	207
7.5	Approaches to Participation Across the Measurement Design Lifecycle	208
7.6	Dimensions of Collaboration in Measurement Design	210
7.7	Methodological Observations & Opportunities	211
	7.7.1 Opportunities to Support Participation Across the Measure- ment Design Lifecycle.	211
	7.7.2 Opportunities to Support Iterative Ideation & Assessment in Measurement Tasks.	214
	7.7.3 Opportunity to Support Larger-scale Participation Without Sac- rificing Expressiveness	217
7.8	Discussion	219
	7.8.1 Limitations	221
7.9	Conclusion	222
	Part 2: Exploratory Case Study on Worker-driven AI Measurement	223
7.10	Introduction	224
7.11	Background	225
	7.11.1 AI Evaluation from a Social Science Measurement Perspective	225
	7.11.2 Supporting Worker Participation in AI Measurement and Eval- uation	226
7.12	Methods	227
	7.12.1 Exploratory Approach and Activities	227

7.12.2	Participant Information	231
7.12.3	Analysis Approach	231
7.12.4	Benchmark Validation	231
7.12.5	Limitations and Ethical Considerations	232
7.13	Designing a Benchmark with Workers	232
7.13.1	Designing realistic test cases	233
7.13.2	Identifying and defining meaningful measurement goals	234
	In the past, workers had tried—but often failed to—prompt the LLM to serve as a reflective aid, rather than a problem-fixer.	234
	By discussing examples of model responses to test cases, work- ers were able to more concretely articulate what ef- fective “people challenging” looks like in their context. . . .	234
7.13.3	Iteratively designing and assessing an LLM-as-a-judge rubric .	235
	Workers primarily focus on designing criteria to <i>measure progress</i> <i>towards positive visions</i> rather than just to <i>avoid unde-</i> <i>sired behaviors</i>	236
	Workers began with an initial five-criterion rubric based on their systematization, and expanded it into a more elaborate rubric through iterative testing and refine- ment	237
	Through the process of operationalization, workers capture important nuances and edge cases beyond their ini- tial concept systematization, drawing on their pro- fessional and lived expertise	237
7.13.4	Rubric Validation	238
	To what extent does the LLM judge align with worker judg- ments?	238
	Can the benchmark differentiate among state-of-the-art base models?	239
7.13.5	Workers’ Experiences from Participation	240
7.14	Discussion	241
7.15	Conclusion	242
8	Concluding Reflections	245
A	Worker-AI Interaction, Part 2	247
A.1	Numbers and percentages of children by race and risk level	247
A.2	Appendix terminology	247
A.3	Comparisons of other disparity metrics and thresholds	248
A.4	Workers see risk labels as screening recommendations	248
A.5	Accuracy breakdown by score	248
B	Improving AI Governance	253
B.1	Deliberation Questions	253
B.1.1	Goals and Intended Use	253
B.1.2	Societal and Legal Considerations	256
B.1.3	Data and Modeling Constraints	259
B.1.4	Organizational Governance Factors	260

C Improving AI Measurement, Part 1	263
C.1 Literature search queries	263
C.2 Papers excluded by criterion	263
C.3 Table of Papers	263
D Improving AI Measurement, Part 2	269
D.1 Additional Methodological Details	269
D.1.1 Phase 1: Selection of use case	269
D.1.2 Phase 2: Supporting workers in creating test cases and anno- tating model responses	269
D.1.3 Phase 2, Phase 3, and Worker-Judge Agreement Validaton: Base models and system instructions for generating model responses to test cases	270
D.1.4 Rubric Validation: System Instructions for Two Conditions . . .	271
D.1.5 Phase 3: Augmenting test cases to support rubric assessment and refinement	273
D.1.6 Additional participant information	273
D.2 Additional Findings	274
D.2.1 Additional examples of rubric refinement process	274
D.2.2 Additional details on workers' experiences	274
D.3 Benchmark Components	275
D.3.1 Test Case Examples	275
D.3.2 Initial Rubric (5 Criteria)	276
D.3.3 Final Rubric (10 Criteria)	277
Bibliography	287

List of Figures

2.1	A high-level overview of the child maltreatment screening process at Allegheny County, illustrating when an ADS assists call screeners' and supervisors' screening decisions.	16
2.2	The current AFST-assisted call screening process. Call screen case-workers make screening recommendations and supervisors make the final screening decisions, both with the AFST's risk score and recommendation.	44
2.3	The interface of the AFST which is used by the child welfare workers to make screening decisions. The left figure shows the score the for a normal referral. The right figure shows a referral with an exceptional high risk that triggers the mandatory screen-in policy.	46
2.4	Black-white screen-in rate disparities for AFST-only and worker-AFST decision	51
2.5	Proportions of Black and white children for which the AFST-only decisions and worker-AFST decisions agreed and differed. Note that the percentages here are over the total number of referrals per race, e.g. the 48.6% in the leftmost bar indicates that 48.6% of all Black children referred to CYF would have been screened in by AFST-only and was screened in by workers aided by the AFST (worker-AFST). Notice that the sum of the two leftmost bars (for a single race) equals the AFST-only screen-in rate —e.g. $48.6\% + 22.3\% = 71\%$ for Black children,— whereas the sum of the leftmost and the right middle bars equals the actual worker-AFST screen-in rate —e.g. $48.6\% + 15.9\% = 61.7\%$	52
2.6	The differences in prediction accuracy between AFST-only and worker-AFST decisions.	59
2.7	A screenshot showing the AFST interface that the call screeners and supervisors see when using the tool [433]. The AFST risk score outputs a single number between 1 (low risk of future placement, leftmost side of bar) and 20 (high risk of future placement, right-most side of bar) for every case. In the image, the tool gave a risk score of 5.	73
2.8	Diagram showing a high-level overview of Allegheny County's child maltreatment screening process, illustrating when an ADS assists call screeners' and supervisors' screening decisions. Taken from [221].	74
2.9	This design concept invited workers to consider not only factors that contributed to an AFST prediction, but also factors that the AFST is unable to take into account.	78
2.10	In this design concept, the worker would be prompted to consider additional factors if their assessment is significantly different from the algorithm's.	78
2.11	This design concept, inspired by worker feedback in the first round, would allow workers to dynamically adjust the factors or individuals that the AFST uses when computing a risk score.	79

2.12	This design concept draws workers' attention to the inherent variability and uncertainty in outcomes predicted by the AFST, using a conventional probability distribution superimposed on the score.	80
2.13	In this design concept, we explore notions of epistemic uncertainty related to the unusualness of a case, and provide feature explanations of why a case could be considered unusual.	80
2.14	This design concept portrays AFST scores as founded on patterns in historical data using observed outcomes from similar cases, including the top three 'nearest neighbors' of the current case.	81
2.15	This design concept maps out a possible way to measure whether a worker's decision is aligned or misaligned with the final outcome, incorporating commonly-used placement and re-referral metrics. . . .	81
2.16	Matrix showing overall perceived helpfulness for each participant (columns) and design concept (rows). As shown by the checkmarks, three design concepts were introduced during the first visit's design interviews. The remaining seven design concepts were introduced during the second visit's design interviews, informed by findings from the field observations and interviews from the first visit.	82
4.1	A visual overview of <i>AI Failure Loops</i> . The six failure modes (in circles) that contribute to the dynamics of the <i>AI Failure Loops</i> exist within a web to illustrate the inter-connected relationship amongst the failure modes.	125
5.1	A high-level overview of the study design. The study included five phases: Onboarding, Pre-Assessment, Training, Post-Assessment, and Post Questions. In the Training phase, participants were randomly assigned between-subjects to the <i>Practice</i> or the <i>Practice + Explicit Feedback</i> case.	144
5.2	A screenshot of the AI Risk Score interface shown to participants in the AI-assisted decision-making task. This mirrors the actual AFST interface used by social workers.	147
5.3	A screenshot of the Case Overview information panel that provides an overview of the alleged child maltreatment case and a text box with the allegations as a qualitative description. The case content, including the allegations, are taken from real historical data recording past experienced workers' AI-assisted decision-making tasks.	149
5.4	A screenshot of the Case Details information panel that provides public records and past referral information on the individuals in the alleged child maltreatment case. The content is taken from the same historical case information used to populate the Case Overview and AI Risk Score.	149
5.5	A screenshot of the question that participants are asked in the AI-assisted decision-making activity (left) and AI score prediction activity (right). To inform this decision, participants also see information from the Case Overview, Case Details, and (for the decision-making task only) the AI prediction.	150

5.6	A screenshot of the six imperfect proxies (screener recommendation, supervisor decision, allegation substantiation, services offered, re-referral, and placement) shown as feedback to participants in the AI-assisted decision-making task in the <i>Practice + Explicit Feedback</i> condition. Participants in the <i>Practice</i> condition were shown the next case immediately after submitting their decision.	152
5.7	A 2x2 confusion matrix breaking down the three alignment measures into four high-level categories.	153
5.8	Mixed effects regression visualizing participants' agreement with the AI prediction, across the 24 training cases. Shaded regions indicate standard error. The purple dotted line indicates average accuracy of the AI model with respect to its (proxy) predictive target.	157
5.9	Mixed effects regression visualizing participants' learning across repeated practice, with respect to the worker-based accuracy measures (i.e., accuracy with respect to past experienced workers).	158
5.10	Participants' Proxy-based Accuracy measures against the baseline accuracy of the model (in purple).	162
6.1	We conducted co-design activities and semi-structured interviews with public sector agency workers (agency leaders, AI practitioners, front-line workers) and community advocates to understand the questions they believed were <i>critical to discuss</i> yet currently overlooked before deciding to move forward with a public sector AI proposal. The <i>Situate AI</i> Guidebook synthesizes these key considerations into a toolkit to scaffold early-stage deliberations around <i>whether</i> and <i>under what conditions</i> to move forward with developing or deploying a proposed public sector AI tool.	168
6.2	Screenshot of a blank board shown to participants at the start of the co-design activity on Mural. This board presents Scenario 1: Discussing ideas for new algorithms to improve services.	175
6.3	Screenshot of a Mural board populated with post-its after one participant's co-design activity. In our iterative co-design process, these post-its were refined by the research team then added to an aggregated list of questions that were successively grouped into higher level categories.	177
6.4	A high-level overview of the main stages involved in the proposed use case for the <i>Situate AI</i> Guidebook, intended to support structured and iterative deliberations within a public sector agency.	186
7.1	The <i>Measurement Design Lifecycle</i> , derived from a synthesis of existing frameworks of both design and measurement.	204
7.2	Heatmap showing the average scores per criterion for each base model using the minimal system instructions.	240
7.3	Heatmap showing the average scores per criterion for each base model using the custom system instructions.	240

A.1	Common group fairness metrics showing disparities between Black and white children. The x-axis is the various decision thresholds for the hypothetical AFST-only decisions (for a given threshold, AFST-only would screen in all referrals with that score or above, and screen out all referrals under). The solid red line represents the worker-AFST final decisions after AFST deployment (2016-2018), the solid blue line with dots represents AFST-only decisions in the same time period (2016-2018). The dashed green line represents the worker-only decisions without before the AFST was deployed (2015-2016), and the dashed orange line with crosses represents the AFST-only decisions generated retrospectively for all referrals in the same time period (2015-2016).	249
A.2	Comparisons of decision outcomes between worker-AFST and AFST-only decisions for referrals between 2016 to 2018. The x-axis is the various decision thresholds for the hypothetical AFST-only decisions (for a given threshold, AFST-only would screen in all referrals with that score or above, and screen out all referrals under).	250
C.1	PRISMA diagram showing how many papers were excluded per criterion.	267

List of Tables

2.1	Participants' self-reported demographics aggregated for call screeners and supervisors.	19
2.2	Total number of children that the AFST-only would have hypothetically screened in and out, compared to the actual decisions made by worker-AFST.	51
2.3	Participants' self-reported demographics aggregated for call screeners and supervisors.	77
4.1	Additional historical and social context on the three feminized occupations presented in our case studies (social work, home health, and K-12 teaching in the United States), drawn from existing academic and grey literature.	127
5.1	Coefficient estimates for models examining pre-post learning effects. Each cell shows the coefficient estimate followed by the standard error for a given term (rows) and model (columns). A positive <i>Phase</i> coefficient indicates an increase in the measure in the post-assessment as compared to the pre-assessment baseline. A visual summary of these findings is reported in Table 5.2.	156
5.2	Visual summary of model findings reported in Table 5.1. Upwards arrow (rose-colored cell) indicates increases in the measure value after the training. Downwards arrow (blue cell) indicates decreases in the measure value after the training, and two downwards arrows (darker blue cell) indicates greater decreases in the measure value in the repeated practice condition compared to when explicit feedback was shown alongside repeated practice.	159
7.1	Metadata about the papers in our corpus.	207
7.2	Descriptive counts comparing the initial versus final rubric, at the start and end of the operationalization phase where workers iteratively refined their rubric.	235
7.3	High-level representation of the final rubric contents, including four main categories synthesized from the criteria. See Appendix D.3, Table D.4 for the full rubric.	236
7.4	Results of the first validation activity showing Pearson's correlation for worker-AI outputs. The cells are color coded as follows: $x \geq 0.9$ is green, $0.9 > x \geq 0.7$ is light green, $0.7 > x \geq 0.4$ is light red, $0.4 > x$ is red.	238
7.5	Average scores (out of five) on workers' LLM-as-a-judge system, for each base model and system instruction condition (customized and minimal system instructions). Delta is difference in average scores between the conditions.	239

A.1	Numbers and proportions of children by race and AFST risk level. Percentages are relative to the total number of children per race. For example, the 9639 Black children labeled mandatory screen in made up 36.9% of all 26123 Black children referred to CYF from August 2016 to May 2018.	247
A.2	Accuracy rates AFST-only and worker-AFST policies binned by AFST score or mandatory screen-in (M) from August 2016 to May 2018 for High risk referrals. The AFST-only policy (with a threshold of 15) would have screened in all referrals in this table, so the accuracy rates in the third column are equivalent to the percent of children who were either re-referred within 2 months or placed within 2 years. For example, of the 15161 children labeled mandatory screen-in (placement model score 18 or higher), 32.1% were re-referred or placed in foster care. The rightmost column is the accuracy of workers' actual screening decisions.	251
A.3	Accuracy rates of AFST-only and worker-AFST policies binned by AFST score and mandatory screen-in (M) from August 2016 to May 2018 for Low risk referrals. Any AFST-only with threshold 10 or above would have screened out all referrals in this table, so the accuracy rates in the first column are equivalent to the percent of children who were neither re-referred within 2 months nor placed within 2 years. For example, of the 81 children with an AFST score of 1, 87.7% were neither re-referred nor placed in foster care. The rightmost column is the accuracy of workers' actual screening decisions.	252
D.1	Overview of participant demographics for the 17 workers who responded to our demographics survey.	273
D.2	The number of workers who participated in different engagements (e.g., workshops) across the three phases, including the length of each workshop.	274
D.3	Initial rubric (5 criteria) developed by workers prior to iterative refinement.	278
D.4	Final rubric (10 criteria) developed by workers through iterative operationalization.	279

Chapter 1

Introduction

AI systems are increasingly used to assist *socially complex* forms of human work, which center on workers' abilities to reason about and respond to the needs, behaviors, and intentions of other people [23, 260, 443]. For example, AI-based tools have been designed to support care workers in providing socio-emotional support [117] or to support social service workers in making decisions about how best to allocate scarce resources [361]. However, research has shown that, in practice, such systems often fail to meaningfully augment and enhance workers' abilities (e.g., [4, 113, 131, 178, 325, 425]). For example, while care robots have been deployed with hopes to improve the efficiency of care work, care professionals have expressed that these robots often increase their workload. Rather than alleviating worker burdens, they required workers to take on additional responsibilities as the maintainers and operators of the robots, contributing to decreased job satisfaction and increased emotional demands [325, 455, 456]. Similarly, in K-12 teaching, the vast majority of technologies have been designed with a focus on supporting student learning, while decentering the needs and practical constraints of the *teachers* who are asked to use them in their classrooms [16, 167, 178, 349]. While policymakers and developers have long lauded these technologies' potential to radically transform K-12 education, researchers have often found that the design of some technologies (e.g., intelligent tutoring systems) conflicts with teacher's pedagogical responsibilities (e.g., maintaining a cohesive learning environment), leading some teachers to abandon them [178, 427].

In my research, I explore how we can design toward more *positive futures for AI and workers* in socially complex work, where AI systems truly enhance workers' expertise and capabilities. My research is grounded in close collaborations with real-world organizations and workers who are either currently using or exploring the use of AI-based tools to support socially complex work. In particular, my work has focused on AI in *feminized labor* contexts such as social work and education, where worker expertise has historically been devalued and misunderstood [114, 125]. Below, I overview two threads of my work: **(1) Understanding** why AI systems often fail to meaningfully support workers in socially complex work (Chapters 2 - 4); and **(2) Designing** new processes and tools to address these challenges (Chapters 5 - 7).

1.1 Understanding why AI systems often fail to meaningfully support workers in socially complex work

In the first thrust of my research, I investigate *why* AI deployments for socially complex work often fail to meaningfully augment worker abilities in practice. I adopt

a social ecological approach [6] to study this question across multiple levels, investigating worker-level, organization-level, and societal-level challenges, and their interplay.

Worker-AI Interaction: How social workers make AI-assisted decisions. I started the first year of my PhD studying how social workers at Allegheny County’s Department of Human Services were using an AI-based risk assessment tool to make child maltreatment screening decisions (Chapter 2). I conducted 37.5 hours of contextual observations and semi-structured interviews with 13 social workers to understand their day-to-day practices and challenges around integrating AI-based predictions into their assessments of child maltreatment referrals (Chapter 2). Overall, while much past research in HCI has emphasized the role of *trust* in shaping workers’ reliance on AI outputs (e.g., [20, 46, 261]), I found that workers’ use of AI predictions in this context was shaped by much more than trust. In particular, it was guided by an interconnected set of model, interface, and organizational factors, from workers’ access to rich, contextual information about the child maltreatment referrals that the AI model did not have access to, and their awareness of misalignments between the AI model’s predictive target versus their own decision-making objectives, to organizational pressures to agree with the AI tool’s output most of the time. While prior reports on the AI tool’s effectiveness and impacts touted improved decision-making performance and positive worker experiences (e.g., [436]), I was surprised to find that workers themselves viewed the AI tool as a missed opportunity to meaningfully augment their own work. Through complementary quantitative analyses of administrative log data on workers’ past AI-assisted decisions, we found further evidence that retaining and augmenting worker expertise in this setting is valuable: By leveraging their understanding of case-specific context to interpret and disagree with potentially biased outputs, workers were able to mitigate potential racial disparities in the AI tool’s outputs (Chapter 2.7). To understand how alternative AI interface and model designs might better augment workers’ practices, we generated a range of design concepts based on our observations and interviews studies (Chapter 2.14.3). While the study led to insights around how different ways of augmenting the AI tool may make the AI tool’s predictions more interpretable to workers, I also observed that many issues workers perceived in the AI tool were reflections of fundamental design flaws (e.g., poor assumptions underlying statistical prediction, like the validity of inference based on historical cases) that could not be addressed without fully redesigning the AI tool.

Overall, my research in Chapter 2 centers *workers’* practices, challenges, and perceptions of the AI tool, and in the process, also sheds light onto how broader organizational factors shape their use of AI. After conducting these studies, one of the questions I felt most pressed to ask required a deeper understanding of these *organizational* factors: How do public sector organizations come to decisions about *whether* and *how* to design and adopt these kinds of AI tools in the first place? While there was a growing body of academic and grey literature documenting cases of AI failure in similar workplace settings, we had minimal understanding of how such foundational decisions were made in the first place.

Organization-AI Interaction: How public sector agencies make AI adoption decisions. To understand *how* decisions to adopt and design AI in high-stakes, complex social decision-making settings (e.g., social work) are made in the first place, in my next project, I adopted the anthropological practice of “studying up” [302] those in

positions of decision-making power at four public sector agencies across the United States (Chapter 3). Through semi-structured interviews and design activities with 16 agency decision-makers (e.g., agency leadership, AI practitioners), I developed an understanding of how decisions to adopt AI are shaped by their *interactions with and assumptions about* other stakeholders within these agencies (e.g., frontline workers), as well as those above that provide policy guidance and resources (legal systems and contracted companies), and those below that they intend to serve (impacted communities). By examining broader power structures and communication barriers across different stakeholder groups, this work surfaced future opportunities for research, design, and policy to support organizations in making better-informed AI adoption decisions.

So far, my research (in Chapter 2 and 3) has focused on AI in human services. As I reflected on my findings in conversations with other researchers who study AI development and deployment in different work contexts, I began to observe an emerging pattern: While many of my findings were strange and unfamiliar to researchers who studied AI systems used by high-status workers such as doctors or judges, researchers who studied AI deployments in occupations like K-12 teaching and home healthcare recognized commonalities with their contexts. This led me to ask: How might broader societal forces impacting workers—in particular, the *historical devaluation of worker expertise* in certain occupations—shape AI design, evaluation, and governance practices, in ways that contribute to AI failures? I explored this question in my next project.

Society-AI Interaction: How societal devaluation of worker expertise shapes AI practices. In my final project under Thread 1, I brought together a team of researchers with experience studying AI development and deployment in a range of feminized labor contexts, following conversations that revealed common, recurring dynamics across the contexts we studied (Chapter 4). Working with this team, I conducted a focused review of academic and grey literature on AI deployments across three feminized labor contexts in the United States (social work, K-12 education, and home healthcare), where work has historically been labeled as “women’s work,” and continues to be predominantly performed by women and people of color typically in low-status positions within organizations [e.g., 77, 83, 84, 414]. This review led to a conceptualization of *AI Failure Loops*: recurring patterns of failure in the design, evaluation, and deployment of workplace AI systems, in which the devaluation of worker expertise leads to ineffective AI deployments—and those failed deployments, in turn, further erode recognition of workers’ skills and expertise. By articulating *why* and *how* AI Failure Loops emerge, I make the argument that workers engaged in devalued labor are especially vulnerable to flawed AI deployments. These AI deployments not only fail to reflect the complexity of workers’ labor, but can also further obscure and erode the visibility of workers’ expertise, creating a negative feedback loop that deepens occupational devaluation. Based on my analysis of past case studies and the concept of AI Failure Loops, I identify opportunities for future research and design to break the loop and avoid these pitfalls.

1.2 Designing processes and tools to engage worker expertise in AI development and deployment decisions

The second thread of my research builds upon the understandings developed in the first thread by demonstrating what alternative approaches to AI development and

deployment might look like, to help avoid such failures. A common finding across my research in prior chapters is that current AI development practices—whether decisions made during problem formulation or measurement decisions made when evaluating an AI system—often reflect misunderstandings of workers’ practices and expertise. This creates barriers to the design and deployment of AI systems that meaningfully augment worker expertise in the context of socially complex work. To address these issues, the second thread of my research focuses on designing novel workflows and processes that can empower workers to shape aspects of AI design, evaluation, and deployment that would otherwise remain opaque and inaccessible to them. I explore how to do this through interventions at multiple points across the AI pipeline: AI use in deployment, AI adoption and governance, and AI measurement.

Improving AI Use: Preparing workers to situate AI predictions against their own complementary knowledge. My research from Chapter 2 showed how workers’ use of complementary information that the AI model could not access was critical in allowing them to make more informed decisions about whether or not to agree with AI model outputs. Workers learned how to combine complementary information sources on their own, through experience making high-stakes, AI-assisted decisions on the job. I wondered whether it would be possible to train workers to make more informed decisions *prior to* using an AI tool in an actual field setting, through low-stakes, simulation-based training. Previous chapters of my research revealed that existing measurements used to evaluate the “accuracy” of AI-assisted worker decision-making (and therefore to evaluate worker “under- or over-reliance” by conventional means) could be misleading due to two assumptions that often do not hold in complex, social decision settings: (1) The availability of a single, reliable “ground truth” label that cleanly maps to both the AI’s predictive target and workers’ decision-making goals [80, 149], and (2) information symmetry between an AI model and workers, where the two have access to the same set of information when making decisions [160, 172]. Therefore, to understand how to train workers towards better calibrating reliance on AI, I first introduce the concept of *critical use*—a process-oriented notion of appropriate reliance that emphasizes workers’ ability to situate AI predictions against potentially complementary knowledge uniquely available to them. I then explore how the design of novel training interfaces can support workers in making more critical use of AI-based risk predictions, using a randomized online experiment with 354 participants. I find that, by providing participants with accelerated, low-stakes opportunities to practice AI-assisted decision-making in this setting, novices came to exhibit patterns of disagreement with AI that resemble those of experienced workers. Overall, our findings surface opportunities for future work to design more effective training interfaces for AI-augmented work in socially complex settings, for example, through the design of improved feedback mechanisms.

Improving AI Governance: Engaging more diverse expertise in AI adoption and governance decisions. As Chapters 2 and 3 describe, many workplace AI failures can be traced back to the earliest stages of design, such as ideation and problem formulation, where inappropriate assumptions are made about workers’ practices and needs, relevant technical and organizational constraints, and other factors. The knowledge and perspectives required to guide more informed assumptions is distributed across different stakeholders—including frontline workers, in-house AI

practitioners, organizational leaders, and impacted community members. However, as discussed in 3, in public sector agencies, most of these stakeholder groups are excluded from formative decision-making. To explore how to support organizations in making better-informed decisions about *whether* and *how* to move forward with a proposed AI tool in the first place, I collaborated with stakeholders across four public sector agencies and three community organizations in the US. Through semi-structured interviews and co-design activities, we developed the *Situate AI Guidebook*—a toolkit designed to bring frontline workers, agency leadership, and AI practitioners together through early-stage deliberations around *whether* and *how* to move forward with a new idea for a public sector AI tool, based on questions around the goals and intended use of the proposed AI tool, societal and legal considerations, data and modeling constraints, and organizational governance factors (Chapter 6). Overall, the resulting set of deliberation questions raised a range of considerations—from centering community needs to addressing potential agency biases, given their positionality—that surface unique considerations for supporting better-informed early-stage design decisions in public sector contexts.

Improving AI Measurement: Enabling workers to shape AI measurement approaches. Across most chapters of my research (Chapters 2– 5), AI challenges in deployment could be traced back to *measurement* assumptions—made when translating from abstract, latent concepts to practical approaches and instruments for measuring those concepts, such as metrics, benchmarks, and LLM-based evaluators [5, 201, 442]. For example, in Chapters 2– 4, I observe how AI-based decision support tools (and their evaluations) in social work are frequently built on invalid measurement assumptions from worker and community perspectives. While social workers aim to assess *immediate safety risks* to a child, AI tools are designed to predict the *long-term risk that a child will be removed* from their home by child services (Chapter 2). These types of measurement issues are pervasive across AI deployment contexts, for example, in social media recommendation platforms [157, 402] and in the design of benchmarks for NLP technologies [268]. The final chapters of my research (Chapter 7) asks the following question: How can we empower non-AI-expert stakeholders, including workers, to shape the measurement decisions underlying AI systems? While exploring related literature on this question, I began to find examples of methodological approaches for stakeholder participation in measurement design (not limited to the context of AI or other technology development) across a range of disciplines—from the social and life sciences, to philosophy of science. My first project in this line of inquiry draws methodological lessons from this broader multidisciplinary literature on participatory measurement design (Chapter 7). Based on a multidisciplinary review of existing approaches, I synthesize a methodological framework for stakeholder participation in measurement design. Through an analysis of trends in existing approaches, I identify opportunities for future research to advance participatory approaches to AI measurement. My second project in this direction empirically explores some of these opportunities in the context of AI measurement for AI-augmented work (Chapter 7.9). In particular, I form a partnership with a local school social work organization to explore how to support *worker-driven AI measurement*—where workers collaboratively shape decisions about *which tasks* the AI system augments, *what* “successful” augmentation means for those tasks, and *how* it should be measured. Through an iterative series of workshops, I engage workers in developing a benchmark (LLM-as-a-judge system and test cases) that captures their understanding of what it means for an AI system to

critically challenge workers' thinking in their specific organizational and work context. Reflecting on our case study, I discuss opportunities for future work to design improved workflows and tools to scaffold worker-driven AI measurement.

1.3 Summary

Overall, my thesis research (1) expands socio-technical understandings of why workplace AI deployments in socially complex work settings often fail to meaningfully augment worker practices. I place this understanding in conversation with prevailing approaches to designing, evaluating, and governing AI systems to (2) identify concrete opportunities to improve AI practices for developers, deployers (e.g., public sector agencies), and users (e.g., frontline workers). And I also address some of these opportunities by demonstrating how (3) new processes and tools can empower worker participation across multiple points of the AI pipeline. In reflecting on my past research and design experiences, I make two main arguments:

Thesis statement

First, it is uniquely challenging to design pro-worker AI for socially complex work. Doing so requires taking a multi-layered approach, by developing an integrative understanding of how AI practices shape and are shaped by individual worker, interpersonal, organizational, and societal systems and structures.

Second, to design towards positive futures for AI in socially complex work, workers themselves must be empowered as expert collaborators in defining what it means and looks like to meaningfully augment their abilities.

Part I

Chapters 2 - 4: Understanding AI Failures

Chapter 2

Worker-AI Interaction

These chapters are taken from the following papers:

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

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

Anna Kawakami*, Venkatesh Sivaraman*, Logan Stapleton, Hao-Fei Cheng, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. ““Why Do I Care What’s Similar?” Probing Challenges in AI-Assisted Child Welfare Decision-making Through Worker-AI Interface Design Concepts.” In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*.

Worker-AI Interaction, Part 1

How Workers Make AI-Assisted Decisions

This chapter is taken from the following paper:

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

Abstract. AI-based decision support tools (ADS) are increasingly used to augment human decision-making in high-stakes, social contexts. As public sector agencies begin to adopt ADS, it is critical that we understand workers' experiences with these systems in practice. In this paper, we present findings from a series of interviews and contextual inquiries at a child welfare agency, to understand how they currently make AI-assisted child maltreatment screening decisions. Overall, we observe how workers' reliance upon the ADS is guided by (1) their knowledge of rich, contextual information beyond what the AI model captures, (2) their beliefs about the ADS's capabilities and limitations relative to their own, (3) organizational pressures and incentives around the use of the ADS, and (4) awareness of misalignments between algorithmic predictions and their own decision-making objectives. Drawing upon these findings, we discuss design implications towards supporting more effective human-AI decision-making.

2.1 Introduction

AI-based decision support tools (ADS) are increasingly used to augment human workers in complex social contexts, including social work, education, healthcare, and criminal justice [95, 147, 169, 182, 261, 459]. These technologies promise to foster better, more equitable decision outcomes by helping to overcome limitations and biases in human judgments [74, 209, 261]. At the same time, AI-based judgments are themselves likely to be imperfect and biased, even if in different ways than humans [95, 226, 261, 439]. Given the potential for humans and AI systems to build upon each others' strengths, and to compensate for each others' limitations, a growing body of research has sought to design for effective *human-AI partnerships*: configurations of humans and AI systems that can draw upon complementary strengths of each [20, 95, 169, 213, 416]. Yet to date, little is known about what factors might foster or hinder effective human-AI partnerships in practice, across different real-world contexts.

In this work, we investigate how social workers at a US-based child welfare agency currently make AI-assisted child maltreatment screening decisions in their day-to-day work. In one form or another, AI-based decision supports are anticipated to play an important role in the future of child welfare decision-making [42, 74, 306, 354, 361]. Yet the use of ADS in child welfare remains contentious, and it is unclear what forms of human-AI partnership might be most effective and appropriate in this context [42, 74, 194, 199, 359]. While some prior work has focused on understanding families' and other affected community members' perspectives and desiderata for the use of ADS in child welfare [42, 64], almost no research has investigated *workers'* experiences working with these systems in practice.

We present findings from a series of interviews and contextual inquiries with child maltreatment hotline workers, aimed at understanding their current practices and challenges in working with an ADS day-to-day. We examine the use of the Allegheny Family Screening Tool (AFST). The AFST was deployed in Allegheny County in 2016, to assist child maltreatment hotline workers in assessing risk and prioritizing among referred cases [74, 436]. The AFST context has been frequently studied in recent years (e.g., [64, 74, 95, 429]), and public sector agencies are beginning to look to the AFST as an example of what AI-assisted decision-making can or should look like in child welfare and similar contexts [354]. However, most prior research on the AFST has relied on retrospective quantitative analyses of workers' decisions, without an understanding of how workers actually integrate the AFST into their decision-making on-the-ground. In this work, we focus on understanding how workers currently use the AFST in their day-to-day work, and what design opportunities exist to support more effective AI-assisted decision-making. We explore the following research questions:

1. How do workers decide when, whether, and how much to rely upon algorithmic recommendations?
2. What limitations and future design opportunities do workers perceive for the AFST or future ADS tools?

We found that, although the AFST had been in use for half a decade, the system remained a source of tension for many workers, who perceived the system's current design as a missed opportunity to effectively complement their own abilities. As a step towards the design of new forms of human-AI partnership in child welfare, we engaged these practitioners in envisioning how future technologies might better

support their needs. In the remainder of this paper, we first provide a brief overview of related work and describe the child welfare decision-making context in which this work is situated. We then describe our contextual inquiries, semi-structured interviews, and analysis approach. Based on our analysis, we present rich findings capturing workers' current practices and challenges in working with the AFST. We discuss how workers' reliance upon the AFST is guided by (1) their knowledge of rich, contextual information beyond what the AI model captures, (2) their beliefs about the ADS's capabilities and limitations relative to their own, (3) organizational pressures and incentives that they perceive around the use of the ADS, and (4) workers' awareness of misalignments between the ADS's predictive targets versus their own decision-making objectives. Based on our findings, we present directions for future research and design implications towards supporting more effective human-AI decision-making in child welfare and beyond. Taken together, this work contributes to ongoing discussions in the literature (e.g., [12, 182, 359]) regarding the need for a broader re-consideration of how ADS should be designed, evaluated, and integrated into public sector contexts.

This work represents the first in-depth qualitative investigation in the literature of workers' current practices and challenges in working with the AFST. Overall, our findings complicate narratives from prior academic and grey literature regarding how the AFST fits into workers' day-to-day decision-making. We expect that the kinds of challenges discussed throughout this paper are not uncommon across AI-assisted public sector decision-making contexts. However, it is uncommon for public sector agencies to open their doors to researchers. We recognize Allegheny County for their strong commitment to transparency, for allowing researchers to closely observe their practices, and for their receptiveness to exploring ways to improve their current practices. We hope that this approach will become a norm in the design, development, and deployment of public sector ADS more broadly.

2.2 Background and Related Work

2.2.1 Designing effective human-AI partnerships

As AI systems are increasingly used to support human work across a range of high-stakes decision making contexts, a growing body of research has sought to design for effective *human-AI partnerships*: configurations of humans and AI systems that draw upon complementary strengths of each [20, 95, 169, 213, 416]. To date, scientific and design knowledge remains scarce regarding what factors foster or hinder effective human-AI partnerships in practice, across different real-world contexts. In some studies, human and machine intelligence combined has been shown to outperform either humans or AI alone [20, 179, 321], while in others human-AI partnerships have failed to improve or have even harmed decision quality, for various reasons [147, 328]. For instance, a long line of literature demonstrates that humans are often either *too skeptical* of useful ADS outputs or *too reliant upon* erroneous or harmfully biased AI outputs (e.g., [47, 102, 147, 258]). Much recent research has utilized large-scale crowd experiments on platforms such as Amazon's Mechanical Turk, to investigate how people integrate algorithmic predictions and recommendations into their decision-making (e.g., [20, 47, 328, 416]). However, experiments with crowdworkers performing simulated tasks may be limited in what they can teach us about decision-making in complex real-world decision contexts such as in social work, criminal justice, education, and healthcare [275, 416].

A complementary line of research in HCI has studied and designed for human–AI partnerships in real-world work settings (e.g., [41, 95, 176, 182, 385, 459]). For example, Yang et al. [459, 461] investigated why efforts to integrate AI-based decision support tools often fail in clinical settings. Building on their findings, the authors designed a radically new form of decision support that better aligned with healthcare workers’ actual needs, routines, and organizational workflows. Similarly, through participatory design workshops with public service (job placement) caseworkers, data scientists, and system developers, Holten Møller et al. [182] found that while data scientists were fixated on developing decision support tools that predicted individual’s risk of long-term unemployment, workers desired fundamentally different forms of decision support. As discussed in the next subsection, relatively few research projects have explored how to effectively design human–AI partnerships for social work contexts [95, 359].

2.2.2 Algorithmic decision support in child welfare

Over the past two decades, many child welfare agencies have begun incorporating computerized decision support tools into various stages of the child protection decision-making process [42, 204]. The most widely-used tools resemble simple checklists, assisting social workers in assessing the risk of a given case based on a small set of manually-entered factors [359, 361]. Yet recent years have seen the introduction of machine learning-based systems such as the privately-owned Rapid Safety Feedback program, deployed in child welfare agencies across several US states [199], and the public Allegheny Family Screening Tool (AFST) in Pennsylvania [436]. These AI-based decision support tools (ADS) make use of hundreds of features that are automatically pulled from multi-system administrative data, promising to support more consistent, fairer decision-making, particularly given that social workers are often operating under limited information and time pressure [74]. However, critics have warned that the use of ADS in child welfare risks amplifying harmful biases in the child welfare system (e.g., reflecting biases in the administrative data on which they are trained and operate upon) [74, 119] or introducing absurd errors into decision processes, which humans would not otherwise make [10, 95, 199].

Several recent attempts to deploy ADS in child welfare have failed due to concerns among affected communities or among the social workers tasked with using these systems (e.g., [199, 359]). To avoid such failures, researchers have highlighted a need for greater community participation during the design phase for new ADS, to ensure that concerns are identified and addressed early on. For example, Brown et al. conducted participatory workshops to better understand the concerns of affected communities, and to identify strategies for improving communities’ comfort with ADS [42]. Similarly, Cheng et al. elicited subjective notions of fairness in child welfare decision-making from parents and social workers across the United States—who did not necessarily have familiarity or experience using existing ADS tools—in order to inform the design of future ADS tools that align with these notions [64].

In contrast to this prior work, the current paper contributes to a nascent body of practitioner-oriented research in HCI that studies the integration of existing ADS in public sector decision-making contexts, with the goal of informing more successful uses of such technologies in the future [92, 261, 359]. For example, drawing upon an ethnographic case study with child welfare caseworkers, Saxena et al. [359] presented a theoretically-derived framework (ADMAPS) to inform the design of algorithmic decision-supports in public sector contexts. The current research takes a

bottom-up approach, validating aspects of the ADMAPS framework, while also presenting several complementary findings. Given that our work considers a machine learning-based system that incorporates hundreds of input variables, in contrast to the simpler algorithmic systems studied by Saxena et al, the intelligibility and transparency of the ADS are key concerns in our context. In another study, De-Arteaga et al. [95] conducted retrospective quantitative analyses of child maltreatment hotline workers' decision-making while using the AFST, finding encouraging evidence that in aggregate, social workers are able to employ their discretion [11, 266] to compensate for erroneous algorithmic risk assessments. We build upon and extend this prior work to gain an on-the-ground understanding of child maltreatment hotline call workers' current practices and challenges in working with the AFST day-to-day, and to gain insight into how they may be able to compensate for algorithmic limitations in practice. Towards the design of new forms of ADS and human-AI partnership in child welfare, we engage these practitioners in envisioning how future technologies might better support their needs.

2.2.3 The Allegheny Family Screening Tool and workplace context

In the United States, the term “child welfare system” refers to a continuum of service: child protection, family preservation, kinship care, foster care placements, and adoption services. The child welfare system's primary purpose is to keep children safe and protect them from harm, activities typically carried out through a series of decisions made in the screening and investigation of abuse or neglect allegations. Its secondary purpose is to connect families to services that will improve conditions in their homes, supporting children at risk. Given the number of children who are reported for maltreatment relative to the number who enter foster care, significant staffing resources are dedicated to the “front-end” of the system: assessments, screenings, and investigations. At the same time, however, most investments have focused on the “back-end” of the system, or those children removed and placed in foster care [156]. This attention has come at the expense of other operational aspects of the child welfare system, such as training and decision support for child maltreatment hotline call screeners and supervisors, caseworkers, or other front-end workers. Today, child welfare agencies in the US are typically staffed and led by social workers who entered the workforce because they were interested in clinical, direct practice with children and families. They do not view themselves as investigators, nor has their professional training necessarily prepared them for the investigative work with which they are tasked [301].

Given the volume of child maltreatment referrals, call screeners and their supervisors struggle to make systematic use of the administrative data and case history available to them. The stakes of the decisions these workers make day-to-day cannot be overstated, as workers face a challenging balancing act between “erring on the side of child safety” versus “erring on the side of family preservation” [74, 364]. As discussed above, in an effort to augment workers' abilities to efficiently process and prioritize among cases, child welfare agencies have begun to turn to new ADS tools. Among the most publicized examples, and the ADS in the context of our study, is the **Allegheny Family Screening Tool (AFST)**. The AFST has been in use at Allegheny County, Pennsylvania's **Office of Children, Youth and Families (CYF)** Intake/Call Screening Department since 2016, where it assists child protection hotline call screeners in assessing risk and prioritizing among referred cases [74, 436]. Figure 2.1 illustrates the CYF screening and investigation process at Allegheny County.

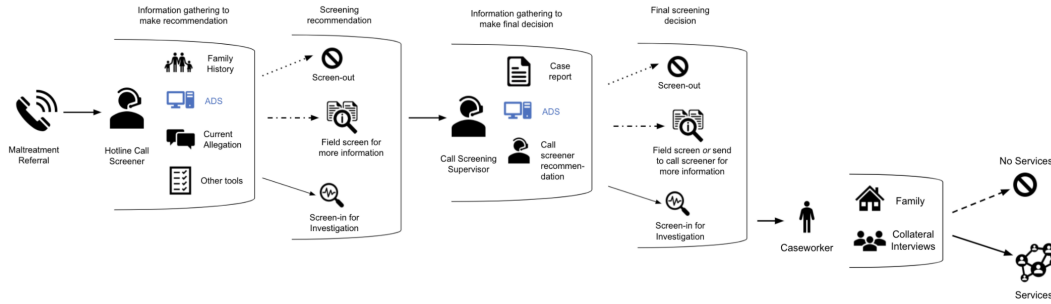


FIGURE 2.1: A high-level overview of the child maltreatment screening process at Allegheny County, illustrating when an ADS assists call screeners’ and supervisors’ screening decisions.

From left to right: an external caller (e.g., a teacher or relative) calls a child maltreatment hotline to make a report, i.e., a **referral**. A call screener is then tasked with recommending whether or not to screen in the report for investigation. The call screener gathers various sources of information to run the AFST, which outputs a score between 1 (low risk of future placement) to 20 (high risk of future placement). Using the AFST score, the current allegation from the caller, and other information sources from public records, the call screener makes a screening recommendation. The call screener may either agree or disagree with a low or high AFST score, when making their recommendation. The supervisor receives the case, along with a case report including the call screener’s recommendation, the AFST score, and other case-related information, to make a final decision. The supervisor may then either agree or disagree with the call screener and/or the AFST score. If the supervisor wishes to screen out a case for which the AFST score is 18 or higher (a **mandatory screen-in score**), they must go through an **override process** to make their screening decision. If the case is screened in instead, it is referred to a caseworker, who might proceed in a number of different directions (e.g., further observation, investigation, or intervention). If the supervisor finds that there is not enough information to make a screening decision, they may send the report back to the call screener to gather more information (e.g., by calling the reporting source).

2.3 Methods

To understand how child maltreatment hotline call screeners and supervisors integrate algorithmic predictions from the AFST into their decision-making, we analyze data from two sources: contextual inquiries and semi-structured interviews with workers. We visited Allegheny County’s CYF where we conducted a total of approximately 37.5 hours of observations and interviews in order to both understand workers’ current decision-making processes using the AFST and to identify design opportunities to support more effective human-algorithm decision-making in the future.

Over the course of two visits, spread across a two-week period, we conducted contextual inquiries and semi-structured post-interviews with a total of nine call screeners and four supervisors. During the contextual inquiries, we observed call screeners taking calls, compiling case reports, running the AFST, and making screening recommendations. We also observed supervisors reviewing call screeners’ recommendations, case reports, and the AFST scores, and then integrating this information to make their final screening decisions. During each visit, members of

our research team conducted observations and interviews with different call screeners and supervisors at the Allegheny CYF office. Each call screener or supervisor was observed and interviewed by one to two researchers, and a different set of call screeners and supervisors was present for each of our visits. Before beginning the observations, we provided the call screener or supervisor an overview of the study purpose and methods and obtained their participation and recording consent. Given that we observed and interviewed call screeners during their working hours, some call screeners or supervisors were unable to complete all study activities. A total of nine call screeners and two supervisors participated in the contextual inquiry (all participants except S3 and S4), and nine call screeners and four supervisors participated in the post-interview. See Table D.1 for aggregated participant demographics. Note that we provide this information in aggregate form to avoid making individual workers identifiable within their workplace.

2.3.1 Contextual inquiries

Playing the role of “apprentices” shadowing a “master” to learn a trade [29], researchers observed call screeners and supervisors as they performed their jobs, asking follow-up questions as needed. For call screeners, this involved observing how they gather information via incoming calls, gather additional information through administrative databases, fill out reports on individual cases, and make screening recommendations with the AFST, amongst other tasks. For supervisors, this involved observing how they made screening decisions using case reports and the AFST score. When making a screening recommendation or decision, the call screeners and supervisors were asked to think aloud [437] to help the researcher follow their thought processes. To ensure minimal disruption, researchers did not ask questions when call screeners were on a call. Researchers asked call screeners and supervisors questions in-the-moment, during their decision-making process (see Supplementary Materials). After making a screening decision, the call screeners and supervisors were asked any relevant follow-up questions in a semi-structured interview style. In total, we observed call screeners taking 32 calls, 7 of which led to making a screening recommendation using the risk assessment tool. The two supervisors were observed for 1.5 hours total and made 14 screening decisions with the risk assessment tool during that time. All call screeners and supervisors consented to having researchers take notes on observations and interviews.

2.3.2 Semi-structured interviews

Towards the end of our contextual inquiries, we conducted a post-interview to (1) validate our observations, (2) gain further insight into call screeners’ and supervisors’ perceptions and practices around the AFST, and (3) understand design opportunities to improve decision-making with the AFST or similar ADS tools in the future. The first iteration of our post-interview protocol consisted of five sections: (1) participant background (e.g., educational and professional background, years of experience working with the AFST, and experience using other algorithmic decision support tools), (2) clarification and validation of findings from our contextual inquiries, (3) understanding worker perceptions of the AFST, (4) understanding worker beliefs and experiences around fairness and bias in the AFST, and (5) understanding where workers perceive opportunities to augment and improve tools like the AFST. After conducting Interpretation Sessions [29] to synthesize findings from our first visit, we iterated on the interview protocol. We added additional questions

to probe deeper on three topics that had come up repeatedly during our first visit: understanding how call screeners and supervisors learn about how the AFST works and behaves, understanding how workers communicate with each other around the AFST's outputs, and understanding worker perceptions of the AFST's effects on overall screen-in rates. See Supplementary Materials for a full list of interview questions participants were asked during each site visit.

All 13 call screeners and supervisors participated (separately) in the post-interview. 12 call screeners and supervisors consented to being audio-recorded and one consented to having notes taken. The average interview time was approximately 47 minutes for the 12 call screeners and supervisors who were audio-recorded. Any call screeners and supervisors from the first visit day who were also present during the second visit day were asked the new questions from the iterated interview protocol. Three call screeners were asked questions from the original interview protocol only, while six call screeners and four supervisors were asked questions from the iterated interview protocol.

2.3.3 Analysis

After each visit, our research team held Interpretation Sessions [29] to collaboratively synthesize findings from our contextual inquiries and post-interviews. Following the two visits, we adopted a thematic analysis approach to analyze approximately 9.5 hours of transcribed audio recordings and 92 pages of notes. Following a shared open coding session to calibrate coding granularity, each participant's data were qualitatively coded by two to three researchers [286]. For each participant's data, we ensured that at least one of the coders was the researcher who conducted the contextual inquiry and/or post-interview with that participant. During this phase, each coder remained open to capturing a broad range of observations in their codes, while also keeping watch for observations related to our original research questions and the six major sections included in the post-interview.

After resolving any disagreements amongst the coders, we conducted a bottom-up affinity diagramming process [29] to iteratively refine and group the resulting 1,529 unique codes into successively higher-level themes. In total, this process generated four levels of themes. The first level clustered our 1,529 codes into 380 themes. These were then clustered into 71 second-level themes, 14 third-level themes, and four fourth-level themes.

The four top-level themes that emerged from this analysis correspond to subsection headers in the Findings section. Each top-level theme captures workers' existing practices with the AFST, as well as design opportunities that they perceive for the AFST to better support their work. For example, within the fourth-level theme on workers' beliefs about the AFST (see Section 2.4.2), there are two third-level themes: One focuses on workers' strategies for gaining greater insight into the AFST's behavior, while the other focuses on characterizing workers' current beliefs about the AFST and the impacts that these have on their day-to-day practice. The former third-level theme includes 11 second-level themes, for example, capturing workers' specific motivations to learn more about the AFST or describing how they collectively make sense of the AFST model through interactions with other workers. Key findings under each of our four top-level themes are presented in the next section.

2.3.4 Ethics and Participant Safety

We assured all workers that their participation was completely voluntary and that their responses would be kept anonymous. To ensure that workers who participated in the study would not be identifiable within their workplace, we only report participant demographics at an aggregate level and omit participant IDs for a small number of sensitive quotes. We acknowledge that HCI projects framed as participatory research and design sometimes take up many hours of participants' time, and then yield no tangible follow-through for participants [326]. We intend to continue our collaboration with workers and other relevant stakeholders in Allegheny County to ensure that they can benefit from this research.

2.3.5 Positionality

We acknowledge that our experiences and positionality shape our perspectives, which guide our research. We are all researchers working in the United States. Our academic backgrounds range across interdisciplinary fields within Computer Science, including HCI and AI. Some of us have prior experiences studying social work contexts or other public-sector decision-making contexts in the United States but not elsewhere in the world. None of us have been investigated by a child welfare agency nor adopted or involved in the foster care system. In addition, none of us have professional experience in child welfare. All authors except two live in Allegheny County; the other two live in Minnesota and California. To conduct this research, we collaborated with Allegheny County's Child, Youth, and Families Department as external researchers. The analysis and writing were conducted independently from the department.

Demographic information	Participant counts
Study participation	Contextual inquiry: All participants except for S3 and S4 Semi-structured interview: All 13 participants
Years in current position	Call screeners: <1 year (1), 1-3 years (3), 3-5 years (1), >= 5 years (4) Supervisors: >= 5 years (4)
Have you ever worked in your current position without the AFST?	Call screeners: Yes (4), No (4), Unsure (1) Supervisors: Yes (3), No (1)
How long have you worked with the AFST?	Call screeners: Since the worker's employment (5), Since the AFST's deployment (3), Unsure (1) Supervisors: Since the AFST's deployment (3), Unsure (1)
Do you have prior experience working in other social work positions?	Call screeners: Caseworker experience (5), Other social work-related experience (2), Undisclosed (2) Supervisors: Caseworker experience (2), Undisclosed (2)

TABLE 2.1: Participants' self-reported demographics aggregated for call screeners and supervisors.

2.4 Results

The AFST has been described as a tool that “*simply augments the human decision whether to investigate a call alleging abuse or neglect*” [67] and “*does not replace clinical judgment but rather provides additional information to assist in the call screening decision making process*” [191]. Meanwhile, HCI researchers have often expressed concern that the presence of tools like the AFST will lead to overreliance by workers who might place too much trust in these tools’ recommendations (e.g., [20, 47, 95, 119, 328]). **Our analysis revealed a more complicated picture that is not fully captured by either of these narratives.**

In this section, we present our findings across four subsections, each of which corresponds to one of the four top-level themes that emerged through our analysis. We first discuss how workers calibrate their reliance on the AFST by drawing upon rich, contextual information beyond what the AI model is able to capture (Section 2.4.1). Then, we discuss how workers’ beliefs about the AFST both shape and are shaped by their day-to-day interactions with the tool (Section 2.4.2). We next discuss the impacts that organizational pressures and incentive structures may have on workers’ reliance, independent of their trust in the technology itself (Section 2.4.3). Finally, we discuss how workers adjust their use of the AFST based on their knowledge of misalignments between the AFST’s prediction task and their own decision-making task as human experts (Section 2.4.4).

Throughout this section, call screeners are identified with a “C,” and supervisors are identified with an “S.” For a small number of sensitive quotes, participant IDs are omitted to provide an additional layer of anonymity.

2.4.1 How workers calibrate their reliance on the AFST based on knowledge beyond the model

In our contextual inquiries and interviews, we observed that workers calibrated their reliance on algorithmic recommendations by **drawing upon their own knowledge of a given referred case, which often complemented the information captured by the AFST model.** Much prior research on human-AI decision-making has focused on settings where the AI system has access to a superset of the task-relevant information available to the human [20, 95, 275]. Yet in many real-world decision-making settings, humans and AI systems may have access to *complementary* sources of information, opening potential for each to help overcome the other’s blindspots [11, 169, 235]. Our qualitative findings validate hypotheses posed in earlier work by De-Arteaga et al. (2020). Through retrospective quantitative analyses of workers’ decision-making with the AFST, these authors found that workers were able to reliably detect and override erroneous AFST recommendations. To explain how workers were able to do this, the authors speculated that the workers may have been cross-checking algorithmic outputs against other relevant information that was available to them but not to the AFST model [95]. Our observations lend credence to this hypothesis: we found that to inform their screening decisions, workers paid close attention to qualitative details which were reflected neither through administrative data nor through manual inputs that workers are able to provide to the AFST.

Based in part on phone conversations with individuals connected to a case, call screeners constructed rich, causal narratives around a given case, accounting for contextual factors that the AFST overlooks such as potential motives of the callers who are filing a report or cultural misunderstandings that may be at play. In turn,

call screeners often communicated such inferences to supervisors, to support their interpretation and decision-making. Observations during our contextual inquiries revealed that, when triangulating across multiple sources of information, call screeners frequently reasoned about possible *causes* behind the evidence they were seeing. For example, as discussed below, workers reasoned about whether a string of re-referrals for a given case (which can increase the AFST risk score [436]) may have been caused by parents repeatedly reporting one another in the midst of a dispute, rather than indicating actual risk to a child. Similarly, workers did not only consider the statistical risk that a factor might pose to a child in theory, as the AFST does, but also reasoned about the plausibility that this risk would be realized in a specific case. For example, in one instance we observed, a call screener saw that a parent had criminal history in the past but a clean record over the past decade. In the context of other indications that the parent had “started a new life,” the call screener judged that the parent’s prior criminal history was not relevant to assessing risk. In addition, if a worker knew that a given parent did not have any contact with a child, they would take this information into account, disregarding potential risk factors tied to that parent. These observations align with recent theoretical frameworks that characterize the role of human discretion in algorithm-assisted public-sector decision-making [11, 359] and describe gaps between human and AI-based decision-making [253]: call screeners were able to reason about potential causal factors based on their on-the-ground knowledge and expertise, complementing the AFST’s use of large-scale statistical patterns.

Both call screeners and supervisors asserted that they decide whether or not to override the AFST primarily based on qualitative details of the allegations, and their understanding of the social context in which those allegations are being made. As S1 put it, the decision of whether to screen-in or override a high AFST score depends entirely on “...*what’s being reported and what the allegations are. You know, and [are] there any safety [concerns] in reference to the younger children in the house?*” (S1). Another supervisor, S2, observed that the AFST is prone to giving high scores for cases that have been frequently referred, and emphasized the importance of considering the allegation information alongside the AFST score, given that the AFST “[*doesn’t consider*] *the whole story*”. Workers shared specific prior experiences in which a high AFST score appeared to reflect a failure to account for relevant context. For example, C7 shared a case in which the AFST score seemed to be high mainly due to the erroneous inclusion of a child’s father in administrative inputs to the AFST model. In this case, the call screener was aware that the father was in jail, and in fact had no contact with the child. Based on this knowledge, and an absence of other evidence of immediate safety concerns, workers chose to override the AFST. In a different case, the AFST provided a low risk score, but workers decided to screen-in the case based on concerning details in the caller’s allegations, pertaining to a relative who was visiting with the child. In this case, the AFST was not aware that the relative was visiting, and was unable to factor in the potential safety risk that the relative posed to the child.

We observed that workers also frequently took into account the social context in which a set of allegations were being made. For instance, when making screening decisions, workers reasoned about the relationship of the caller making an allegation to the child, the family, and the alleged perpetrators, as well as the caller’s cultural background and potential motivations. During our contextual inquiries, a few call screeners encountered cases that they believed had high AFST scores mainly because these cases had been re-referred several times, not because the child was truly in any danger. Rather, based on their phone conversations with individuals connected to a

case, these call screeners believed that these multiple re-referred cases represented “retaliation reports” (e.g., parents repeatedly reporting one another in the midst of a dispute) or cross-cultural misunderstandings on the part of the caller (see Section 2.4.2).

In contrast to workers’ focus on details of the allegations, the AFST is only able to account for the current allegations at a coarse-grained level. The AFST factors in a categorical “allegation type” variable, taking into account the broad types of allegations made in both current and past referrals. For example, workers can select whether the allegation in a given case maps to categories such as “Child Behaviors,” “Caregiver Substance Abuse,” “No/Inadequate Home,” “Neglect,” “Physical Altercation,” or “Parent/Child Conflict” [436]. Workers perceived allegations related to “Imminent Risks” or “Caregiver Substance Abuse” as posing immediate safety risks to the child, and therefore strong indicators of potential child maltreatment; on the other hand, they viewed allegations related to “Parent/Child Conflict” as requiring more context, given that not all forms of conflict between parents and their children are necessarily cause for alarm, and cross-cultural differences may influence how conflicts are perceived.

The inclusion of categorical variables that capture aspects of the allegations was added to the AFST in 2018, in response to workers’ desires to have allegation information factor into the AFST’s risk score calculation [436]. **However, given a fixed set of allegation types available for workers to select, workers are currently limited in their ability to communicate relevant contextual knowledge to the model via this mechanism.** As such, the inclusion of “allegation type” variables does not eliminate the need for workers to calibrate their reliance on the model based upon their personal knowledge of a case. Despite these limitations, it is worth noting that allegation type is currently one of the few places in a referral where call screeners have discretion to influence the risk assessment score based on their own interpretations. Workers desired the ability to better incorporate their knowledge of relevant context into the score calculation. For example, C3 proposed that it would be helpful if they could leverage their knowledge of case-specific context to explore counterfactuals, by “[removing] some things out of the score [... to] kinda play with it and say, ‘if you take this out of there, what kind of score would this person get?’ ”

We next discuss how workers’ perceptions of the AFST’s capabilities and limitations, including its inability to account for many qualitative details of individual referrals, impact how they work with the AFST.

2.4.2 How workers’ beliefs about the AFST shape and are shaped by their day-to-day use of the algorithm, in the absence of detailed training

Given limited formal opportunities to learn how the AFST works, we found that social workers improvised ways to learn about the AFST’s capabilities and limitations themselves—both individually, through their day-to-day use of the tool, and collectively, by sharing their observations and inferences about the tool between workers. In turn, the beliefs that workers develop about the AFST play a major role in shaping how they work with (and around) the algorithm. Below, we first describe how workers attempt to learn about the AFST’s capabilities and limitations, and discuss their desires for opportunities to learn more about how the AFST works. We then present aspects of workers’ mental models of the AFST—i.e., the internal representations or systems of beliefs that they build up through experience, which they use to understand, explain, and predict algorithmic behavior, and then act accordingly [18,

36, 236, 242]—and we discuss the impacts that these beliefs have on their decision-making and communicative practices around the AFST.

How workers (try to) learn about the AFST’s capabilities and limitations through day-to-day use of the tool

Despite having used the AFST day-to-day for multiple years, we found that most call screeners and supervisors know very little about how the AFST works or what data it relies upon. In one instance, a call screener (C3) turned to the interviewer to ask if they knew how the AFST worked, hoping that they could answer their longstanding questions about what features the AFST uses. Workers described receiving surface-level procedural training on how to run and view the AFST score a number of years prior (S1, S2, S3, C3, C8, C2), but few could recall being presented with authoritative information on details such as the predictive features used by the AFST, or how these features were weighted. Moreover, workers’ understanding is not aided by the AFST interface, which simply provides a numerical risk score or a mandatory screen-in/out message.

While Allegheny County is outwardly transparent about the AFST’s design and development process, with public-facing materials and documentation published on the web (e.g., [189, 436]), **workers are intentionally provided with minimal information about the AFST model to discourage ‘gaming the system’ behavior.** That is, the designers and administrators of the AFST had worried that if too much transparency were provided to workers, this might enable workers to strategically manipulate data inputs used by the algorithm, in order to get a desired output. While this is a real concern in algorithm-assisted child welfare decision-making contexts [359], we observed that **given minimal information about the AFST’s capabilities, limitations, and overall functioning, workers took matters into their own hands, improvising ways to learn more about the tool themselves.**

Through their day-to-day interactions with the AFST, workers built up intuitions about the AFST’s behavior and hypotheses about its limitations and biases. For example, **call screeners described a collaborative, game-like approach to predicting AFST scores**, recalling past occasions where call screeners would informally “gamble on what this score’s gonna be” (C7).

“What we do is we have to research everything and... start up a report. And then I see, ‘Oh, wow. They have a lot of history. Oh, they’ve got this going on and they have that going on...’ So then I’ll say, I bet this score’s gonna be...”
(C3)

In addition to honing their ability to predict AFST scores through these guessing games, **workers would sometimes discuss the AFST score for a given case amongst themselves, in order to collaboratively make sense of it.** Workers did so informally, even though the screening protocol officially assigns only one call screener and one supervisor per case. Describing their open workspace and collaborative decision-making process, C8 said “I hear everything [...] we’re sitting on each other’s laps”. Similarly, S1 explained that “it’s not like we all work in a little bubble [...] If we’re struggling with something, we will [...] have a little, brief discussion [...] We say, ‘Well hey, this score is saying it’s a 19, but we don’t really see anything on this page based on these allegations that says it should be a 19.’”

Workers also developed strategies to gain a more direct window into the AFST's behavior. For example, in order to learn more about the impacts that particular factors have on the AFST score, both call screeners and supervisors described **computing scores on slightly different versions of the same referral data and then comparing them to draw inferences**. Given that the AFST reads from saved referral records, workers cannot significantly modify its inputs without sacrificing thoroughness or accuracy of documentation. However, they can still gain some insight into the impacts that particular factors may have on the score by making small adjustments. For example, to understand the impact that a particular family member's administrative records have on the AFST score, a worker might omit that family member from the AFST score calculation, run the AFST to see the score, and then compute another score with the family member included and compare the two scores.

Workers are frequently surprised and confused by the AFST's behavior, and want support in learning about the tool

Echoing prior findings from Eubanks [119], these experiences using and tinkering with the tool enabled workers to make predictions of the AFST score that were often accurate. Yet importantly, the intuitions that workers had developed around the AFST were imperfect: call screeners and supervisors **often encountered AFST scores that surprised them, reminding them that their understanding of the tool's behavior is limited**. In reality, the AFST model makes predictions based on a wide array of factors, and C6 lamented that call screeners can *"constantly get into a place where we're assuming that [one particular factor] is what has to be driving the score."* Workers shared specific instances in which their assumptions were violated. For example, referencing a belief that the presence of many historical referrals and service engagement leads to higher AFST scores (see Section 2.4.2), C2 acknowledged that *"just because you have a long history doesn't necessarily mean that you're going to get a big score. So we just guess that that's potentially what it could be. I'm sure there are other factors... but we don't know for sure."* In other cases, workers found that their assumptions about the model were violated when they observed the AFST score shifting in response to changes in data fields that they had not previously thought the AFST took into account (C7, S3). For example, during our observation, C7 stated that the allegations have no impact on the AFST score. However, when rerunning the AFST for an existing case with a previous "Truancy" allegation and newly added current allegations on "Caregiver Substance Abuse" and "Neglect," the call screener observed the score increasing.

While workers sometimes used unexpected scores to adjust their understanding of the model, at other times, they were simply confused when they could not identify potential causes for unexpected behaviors. For example, one supervisor (S2) noticed the score changing seemingly at random when they ran it multiple times, even from a 14 (moderate risk) to 7 (low risk), with no apparent changes to the algorithm's inputs. Regardless of what factors led to unexpected model behavior, **these experiences led some workers to perceive the AFST as hyper-sensitive or unreliable, contributing to their overall distrust of the tool** (C1, C7, S2). C7 compared the expectation that workers should use the AFST, despite lack of insight into how it works, to blind faith: *"not knowing what data's going into it [...] it's more of having, like, religious faith, I guess, you know?"*

Given these challenges, workers desired more opportunities to learn how the AFST works, to empower them to work with the tool more effectively, and to help them avoid making faulty assumptions about AFST scores in particular instances.

As discussed in Section 2.4.3, while the AFST includes a field in which workers can submit feedback and questions to the AFST maintainers, workers expressed that they have rarely received responses that they personally found useful and understandable. Several workers believed that having greater insight into how the model works would enable them to better judge when they can trust the score, and ultimately allow them to make better use of the AFST (C3, C6, C7, C9, S3). As C3 put it, *“If we knew more about how we got to the score, I think I’d pay more attention to how the score is going to help me”*. One call screener (C6) expressed that they saw greater transparency around the AFST score calculation as an opportunity to balance out power imbalances: *“just knowing these are the top three things influencing the score or, you know, something to that effect. Yeah, I mean, it feels like there are ways that it could be better integrated into the workflow [to] make it feel collaborative versus telling me what to do”* (C6).

At the time of this study, workers were interested in learning more about how the AFST produces its risk assessment score at nearly every stage of the model’s pipeline. For instance, although workers were aware that the AFST relies upon administrative data that they can view through their web interface, they wanted to know whether it also draws from other county databases, accounting for data that they themselves cannot easily access (C2, C8, S3). Many workers also desired more opportunities to learn about the specific factors that the AFST is using, and how those factors are weighted (C2, C3, C6, C7, C8, S1, S2, S3). For example, workers were interested in learning which individuals in the referral contributed to the score and by how much (C3, C6, C8, S3), even wondering whether it was possible for deceased relatives to affect the AFST’s assessment of risk to a child (C8). Beyond new forms of training, several workers expressed desires for new decision-time interfaces that could support their current, informal practices for learning about the AFST’s behavior. For example, an augmented version of the AFST interface might explicitly support workers in rapidly exploring multiple counterfactual inputs to the AFST (cf. [449]) in order to learn how particular factors impact the AFST score in specific cases (C1, C6, C8, S1).

Workers’ beliefs about the factors used by AFST influence their reliance on the score

Workers bring a range of knowledge and beliefs with them to the task of interpreting an AFST score and integrating it into their decision-making, including both their knowledge about the broader context of a case (Section 2.4.1) and their beliefs about how the AFST score is computed (Section 2.4.2). In our contextual inquiries and interviews, most workers referenced beliefs about how the AFST considered the following four factors, among others: the number of re-referrals on a case, the extent of a family’s prior involvement with public services, the size of the family, and the age of the alleged victim(s). During our observations, **workers often made guesses about how each of these features might be influencing the AFST score calculation in a given case**, and they reasoned about whether or not a feature’s potential influence on the score was appropriate. **These inferences, in turn, informed their perceptions of whether the score was “too high” or “too low” for a given case:** *“the more you use [the AFST], you kind of pick up why it will go a certain way and you can kind of use that in addition to what you do know to make an appropriate determination”* (C1).

Many workers believed that a case with **high numbers of re-referrals** consistently resulted in higher scores (C2, C4, C7, C9, S2). This aligns with the AFST’s use of past referrals as a positively weighted predictive feature [436]. To illustrate, C2

described prior experiences with “retaliation reports” in which parents in custody battles call on each other repeatedly, generating enough referrals that *“eventually, the score is going to be ‘Accept [for investigation]’ regardless of what [the referral] is.”* Workers often used their knowledge of the allegation details to override high AFST scores that appeared to be driven by high numbers of re-referrals (see Section 2.4.1). While some workers understood that the **“type of allegation”** factored into the AFST score (C1, S3), others were uncertain or believed that it did not have much of an impact (C9, S1, S2). In reality, as discussed in Section 2.4.1, since 2018, the AFST model has factored in some information about the allegation through a categorical variable that captures broad categories of allegations for current and past referrals. Workers’ beliefs about the impact of allegation information on the AFST score appeared to influence their use of the tool. For example, one participant, who believed that allegation type impacted the AFST score, mentioned not wanting *“to go super crazy on allegations”* when entering data on the allegation type, if they thought it would unfairly drive up the AFST score.

Workers also believed that the greater a **family’s degree of involvement with the “system” and “services”**—including public mental health and other medical treatment, criminal history, and welfare records—the greater the AFST score would be. In reality, the AFST model does use behavioral health records (with a small positive weight for substance abuse, but a small negative weight for neurotic disorders) as well as criminal history (with positive weights for features related to whether the victim is in juvenile probation currently or in the past year). However, the AFST no longer uses public welfare records as of November 2018 [436]. When reviewing individual cases, workers often justified their disagreement with the AFST score with reference to their beliefs about the AFST’S use of system involvement as part of its score calculation. For instance, C9 explained a high AFST score for a family with no referral history by pointing out the father’s criminal record and mother’s drug use. Describing the limitations with quantitatively assessing system involvement as a predictor for child maltreatment, S1 said, *“If I’m someone who accesses the system, meaning... I get public assistance, I’ve committed a crime... all that stuff stays in the system, and that’s the driving factor for that [AFST] score, but that doesn’t mean that the subject child in that report is being maltreated or being in danger.”*

Similarly, workers believed that **reports with more people in them** received higher AFST scores. Workers noticed the impact that the number of people had on the AFST score, both within and across individual case reports. For example, C5 said, *“the more people that are involved with these families, no matter what it’s for, the higher their score’s gonna be.”* Similarly, C2 described that they often see a case’s AFST’s score increase when they add additional people to the report: *“We had one report where the mom... never had CYF history ever before. Dad was added to the report: [the AFST’s recommendation] was automatic screen-in because he has kids with other people who have a long history with us.”* The belief that reports with more people yield broader AFST scores broadly aligns with the AFST model’s use of the number of children, perpetrators, parents, and victims as predictive features [436]. However, although the numbers of perpetrators, parents, and victims have positive feature weights, the number of children (of all age groups, from 3 to 18) actually has a negative feature weight.

Given that the AFST does not account for the type of relationship the perpetrator or parent has with the victim, workers expressed concern that when more individuals (along with their respective histories of system involvement) are added to the report, the score may shift in ways that do not reflect actual changes to the child’s actual safety risk. C8 described that workers sometimes have enough contextual

information about the individual's relationship to the victim to assess whether increases in the AFST score are justified. Other times, however, they may not know enough and are left wondering whether the child is truly more at risk because of the added individual:

"I feel like everybody's got a number, all right? [...] The problem is [all the individuals added to the report are] hooked with [the victim. . .] And I wonder sometimes to what degree do people who aren't actually living in the house and affecting the children affect the score. [...] the thing is when you're running the score here, we don't know that, you know, Dad hasn't seen little Deon since he was 3 and he's now 15. [...But] some of it you would know" (C8).

Similarly, when explaining why they disagreed with the AFST score's use of number of people as a predictive feature, C3 related the potential impact to her personal life: *"I have a niece who has three children and three fathers and she's really a good mother. And if a call came in on her, I'm sure it's gonna be a high score [...] that's the part I think is unfair."*

Finally, workers believed that **younger alleged victims** received higher AFST scores. This is largely correct, as the AFST model weights infant, toddler, and preschool victims more heavily than teenage victims [436]. Workers were also aware of a protocol in which the AFST automatically screens in high-risk referrals involving children under the age of three. Unlike the previous factors, this one tended to align with workers' own assessments: in general, workers agreed that younger victims were at greater risk than older ones (C1, C3, S2). However, workers also perceived that **older alleged victims**, especially teenage victims, were sometimes given unjustifiably high or low AFST scores, depending on if the person had a history of system involvement. As a result, workers pay close attention to the reasons for referral associated with the case, alongside the alleged victim's history of system involvement, in order to inform or justify decisions to disagree with the AFST:

"[It is easy to identify] when the score will appear wrong [...] when we have reports of kids that are truant. When we get a high-risk score for a kid, the only reason why the report is being made is because the kid is truant, and truancy doesn't put a 15 or 16-year-old kid at high risk. And, you know, it's based on them pulling information from this program, this program, this program, and that program, [which] does not impact [whether] the kid is going to school or not." (S1).

Beyond the features described above, some workers expressed confusion at how the model factors in other features. For example, workers perceived that the AFST tended to assign higher risk scores to families from underprivileged racial identities and socioeconomic backgrounds (C1, C2, S2). However, they were unsure whether to attribute high AFST scores to race and socioeconomic status directly, or to other correlated features (C3, C7, C9, S1). For example, while discussing the compounding effects of race and income, S1 said that if the AFST gives a high score: *"[I speculate it must be from disparity because if] I'm not getting money from [public assistance] the computer can't judge me." (S1).*

Overall, workers' intuitions about the model reflected predictive features that are actually used by the AFST model. However, given minimal transparency around the model, workers sometimes **inferred overly general patterns based on their observations** of the AFST's behavior, or **misjudged the direction of a feature's influence**. Furthermore, the accuracy of workers' beliefs about the magnitude of particular features' influence remains unclear.

2.4.3 Beyond “trust” in AI: Influences of organizational pressures and incentives

Prior work has often understood human reliance on algorithmic decision supports in terms of patterns of “over-reliance,” “under-reliance,” or “appropriate reliance”, that are shaped according to how well-calibrated an individual’s trust is in a given algorithmic system [20, 47, 258, 328]. However, this conceptualization overlooks the potential impacts that organizational pressures and incentive structures may have on workers’ reliance, independent of their trust in the technology itself. In this section, we first describe the impacts of perceived organizational pressures upon workers’ reliance on the AFST: in many cases, workers agreed with the AFST score not because they saw value in the tool or because they trusted the technology, but rather because they perceived organizational pressures to do so. We then discuss the influences of organizational incentives on workers’ motivations to disagree with algorithmic recommendations or to provide feedback towards improving the AFST model.

Organizational performance measures shape workers’ reliance on algorithmic recommendations

Several call screeners and supervisors perceived **organizational pressures to avoid disagreeing with the AFST score “too often.”** Workers had either heard about or attended monthly meetings where an internal analytics team discusses how often the workers override mandatory screen-in protocols. Workers shared that while they do not know for sure whether the analytics team has a specific “acceptable” rate of overrides, they perceived that in some months, they had crossed an unspoken line by overriding too often. As one supervisor said, *“someone’s watching to make sure that the majority of those high risk protocols are being assigned.”* **Given these perceived pressures, some call screeners and supervisors said that they will sometimes agree with the AFST, going against their own best judgment:**

“It’s not uncommon to be faced with a higher score, but the allegation that is contained in the report is, like, a low risk kind of allegation, you know? [...] The score is high risk protocol because of all of the other stuff that the tool runs, the algorithm gets you a high score. [...] But] there’s no safety concerns at this point. So then the screener will say, do I have to assign this? Because it says high risk protocol. Do I have to? So I’ll help walk through it [with them] because you don’t want to be accused of not using the tool, because there is regular oversight from [the] analytics team.”

Even when workers did not feel personally affected by these pressures or incentives, they took them into account when reasoning about others’ motivations. For example, when reflecting on what would motivate supervisors to disagree with call screeners’ recommendations and align with the AFST score instead, a call screener guessed that it may be to *“align with directives from administration.”*

Although some workers felt pressure to make decisions against their own judgement, they expressed awareness and concern regarding the impacts their decisions could have on both families and caseworkers. Workers believed that, if they were to always agree with the AFST, they would have many more screen-in decisions than necessary. When describing the extent to which the AFST may increase screen-in rates, C5 explained *“I think that if we [only followed the AFST], we’d have to hire at least 50% more [caseworkers] than we have now.”* This led some workers to worry about

the impacts increased screen-in decisions might have on caseworkers, who must follow up with an investigation on all cases that are screened in. For example, during our observations of S2, this supervisor decided to screen out a case with a middle-range AFST score of 13, reasoning that they should avoid burdening caseworkers. S2 explained that caseworkers have sometimes complained to them about their decisions, in cases where they screened in high risk protocol cases due to organizational pressures, against their own judgment. They expressed that the situation is “tricky” because they cannot easily explain their decision to the caseworker, given that caseworkers do not see the AFST score. **Caught between conflicting pressures from the administration and from caseworkers, some workers wished the AFST tool would provide not just a number, but support in justifying their decision to those whom it would affect (C6, C7, C9, S2).**

“The input does not feel like a two-way street:” Workers’ motivations to disagree with or provide feedback on algorithmic recommendations are shaped by organizational incentives

Workers’ decisions to agree with a high-risk protocol AFST score, even in cases where they disagree, may sometimes be influenced by **a desire to avoid additional work that they perceive as unnecessary**. When deciding to override a high-risk protocol, workers are required to write open-text responses describing their rationale behind the override—a step that the AFST’s designers intended to produce friction and promote worker reflection [436]. However, workers did not see value in this override process, which they found tedious.

Moreover, although the open text field was partly designed to give workers an opportunity to provide feedback on potentially erroneous AFST scores, **workers were unsure whether or how their feedback would actually be used**. For instance, one supervisor expressed discontent that they usually did not receive any acknowledgement that their feedback had been read and considered, contributing to their feeling that writing these “narratives” was a waste of their time, which unnecessarily increased their workload. Another participant simply said “there’s no point” in providing feedback on the AFST.

When workers did receive responses to feedback, **they perceived that the administration tended to discount feedback regarding potential limitations and improvements to the AFST**. This left workers feeling that their own expertise, as human decision makers, was being undermined while the effectiveness of the AFST was being defended. Both call screeners and supervisors said that they had received responses that they perceived as unhelpful, dismissive, or even hostile, leading some to “give up” on giving feedback:

“The input does not feel like a two-way street. It’s, like, we are told why we’re wrong. And we just don’t actually understand algorithms versus, like, maybe you [the workers] are observing something that [our algorithm] missed, you know? [...] in the past, it’s felt like when we’ve taken advantage of that function [to give feedback on the AFST score], it has kind of resulted in just explaining to you why the score is actually right. [...] I think it’s kind of been a point of conflict in the past.”

Given these experiences, workers shared that they have critical discussions about the AFST’s behavior with their peers less often than they used to. In addition, workers believed that they are now less likely to disagree with the AFST than they once

were, having acclimated to using the tool and **having given up on the idea that their feedback can contribute meaningfully to improving the tool’s usefulness.**

So far, we have described how workers’ trust and reliance upon the AFST is influenced by organizational decisions relating to model transparency, choices of performance measures, and processes for overriding or providing feedback on algorithmic recommendations. In the next section, we describe how decisions about the design of the AFST model itself influenced workers’ perceptions and practices with the tool.

2.4.4 Navigating value misalignments between algorithmic predictions and human decisions

Workers often described differences between the criteria that they personally use to make decisions (ensuring that children are safe in the near term) versus the targets that the AFST predicts (risk of particular adverse outcomes over a two-year timespan). Due to these misalignments, several workers did not view the current version of the AFST as particularly relevant to the decisions they need to make day-to-day.

Workers’ awareness of misalignments between the AFST’s predictive targets and their own decision targets

Workers’ focus on immediate safety and short-term risk has long been a point of contention with administration at the Allegheny DHS, and it has often been attributed to a lack of trust in the AFST. For example, in a 2018 interview, a DHS deputy director expressed that screeners *“want to focus on the immediate allegation, not the child’s future risk a year or two down the line [...] Getting them to trust that a score on a computer screen is telling them something real is a process”* [194]. While this narrative aligns with some of our findings, we found that **workers’ concerns around these misalignments extend well beyond a lack of trust in the underlying technology.** Some workers disagreed with the very idea of making screening decisions based on predictions of longer-term risk, perceiving this problem formulation as **fundamentally misaligned with their roles and responsibilities.** Other workers viewed the AFST’s focus on longer-term outcomes as **complementary to their own focus on immediate safety concerns.** However, while these workers were open, in theory, to the idea of factoring in longer-term risk into their decision-making, **they felt confused about how exactly they were expected to do so in practice.**

Both supervisors and call screeners described that they make decisions based on evidence of immediate risks to a child’s safety, in contrast to the algorithm’s focus on predicting longer-term risk. One supervisor, S2, worried that call screeners may interpret the AFST score as predicting “risk” without considering that its definition of risk does not necessarily imply immediate safety risk to a child. However, call screeners exhibited some awareness of the AFST’s notion of risk. For example, one call screener noted that:

“There are times where [knowing] the risk of removal within two years is not feeling like it’s super relevant to the decision that is needed. And it [has] very little to do with immediate safety or anything like that.”

Most workers expressed that the AFST plays a relatively minor, non-driving role in their decision-making processes, overall. When explaining why they avoid relying on the algorithm, these workers invoked not only the AFST’s perceived limitations (described in Section 2.4.2), but also concerns about the outcomes that the

tool predicts. For example, C6 shared that they sometimes feel that their personal goals of keeping children safe clash with the algorithm's prediction targets: *"it just has always felt like the risk of removal in two years is inherently going to be increased by our involvement, because we're the only ones that can remove the children."* They also worried about possible contradictions between the advertised goals of the AFST versus its use of proxy targets such as re-referral and placement:

"To me, the tool has always been described as something to identify the families that might typically slip through the cracks or that wouldn't be most obvious upon initial assessment [...] Like, those families [that are frequently investigated because of the AFST every time they are referred] are not [the] ones that are slipping through the cracks and they're tripping on every single crack that they seem to encounter." (C6)

Some workers went so far as to claim that they *never* change their decisions based on the AFST score. For example, C4 said, *"I look at the score. I often, you know, am in agreement with it. I think it, you know, does a good job trying to pull everything together and come up with the best possible solution with the score. [...] No [the AFST score does not change my recommendation]. I decide whatever I think."* C2 simply said, *"I hate it [...] I don't think it should have a role, period, honestly"*. Others, like C6, described that when making recommendations, *"[The AFST] doesn't feel like much more than a nudge in one direction or the other."* With that said, **observations during our contextual inquiries indicate that some of these participants may be underestimating the influence the AFST has on their decisions.** For example, C1 had claimed that *"I personally, haven't had an instance where it caused me to change my recommendation."* However, in alignment with findings discussed in Section 2.4.3 regarding the impacts of perceived organizational pressures on workers' practices, during a contextual inquiry, the worker was observed changing their screening recommendation from screen-out to screen-in after seeing the AFST score on a referral, explaining *"I have to recommend it"* because the AFST score was high (even though it was not high enough to require a mandatory screen-in).

Worker beliefs about complementarity between the AFST and themselves

Despite their awareness of misalignments between what the AFST predicts and the actual decisions they are trying to make, some workers saw value in using the AFST. Although workers were aware of potential biases and limitations of the AFST, **some believed that using the AFST might help to mitigate some of their personal biases.** For example, C6 believed that the AFST score may help them mitigate their personal biases arising from cultural differences between themselves and the family:

"I think [...] a dirty house or something like that, I feel like those are often the ones where you're not sure if it is just a value or moral judgment and so, you know, the tool can probably weigh a little bit more in those circumstances [...] for those kinds of situations where, like, if I know that my personal bias may be underplaying a risk or something, [the AFST score] can be helpful." (C6)

Similarly, C9 reflected that *"prior to the algorithm, things were screened out or not screened out [by the supervisor] and I would be surprised by that decision. I think, like, we're being a little more consistent with our decisions [now]."*

Despite several workers describing clear value misalignments with the algorithm's prediction targets, all supervisors and some call screeners (C5, C6, C7, C9) said that they **use the AFST score to help them make screening decisions in cases**

where they personally feel uncertain about a decision or, for one participant, when they are pressed for time. For example, S3 described that, for cases where they are initially uncertain about what screening decision to make, seeing the AFST score can help them more confidently make a decision: *"Maybe it's something iffy, and then I run the score, and then the score is high. And then it's like, 'Oh, yeah, definitely it should be assigned.'"* Another supervisor, S2, agreed that the AFST score is helpful when the cases are not *"straightforward,"* providing physical discipline of a child as an example. Similarly, C6 said that they find the AFST score most helpful when they are *"on the fence"* about a decision. C9 noted that they usually follow the AFST's recommendation more *"if you're in a hurry."*

That said, workers expressed that they have a harder time interpreting and using the AFST score when the score falls in the middle range (e.g., between 10 and 14), perceiving these cases as missed opportunities for the AFST to effectively complement their own judgment. C5 explained that *"that's where [supervisors] have the most problem, too [...they'll] say every once in a while, like, 'You got me [going] crazy over these f**ing yellow reports!'"* Given minimal transparency into what factors contribute to a mid-range score, and no information from the AFST beyond the score itself (Section 2.4.2), workers generally disregarded the AFST's outputs in these mid-range cases. **Workers expressed a desire to have the AFST communicate back, not "just a number" but also additional context explaining the AFST score,** for example, to assist them in interpreting middle range scores and in integrating the score into their decision-making.

2.5 Discussion

As child welfare agencies increasingly adopt ADS to assist social workers' day-to-day work [42, 74, 306, 354, 361], it is critical that we understand workers' experiences with these systems in practice. In this paper, we conducted the first in-depth qualitative investigation in the literature of child welfare workers' current practices and challenges in working with a prominent ADS (the AFST) day-to-day. The AFST context has been frequently studied in recent years (e.g., [64, 74, 95, 429]), and public sector agencies are beginning to look to the AFST as an example of what AI-assisted decision-making can or should look like in child welfare and similar contexts [354]. However, most prior research on the AFST has relied on retrospective quantitative analyses of workers' decisions, without an understanding of how workers actually integrate the AFST into their decision-making on-the-ground.

Through a series of contextual inquiries and semi-structured interviews, we observed ways in which workers' reliance and trust on the AFST are guided by (1) their knowledge of rich, contextual information beyond what the underlying AI model captures, (2) their beliefs about the AFST's capabilities and limitations in relation to their own, (3) organizational pressures and incentives that they perceive around the use of the ADS, and (4) their awareness of misalignments between the ADS's predictive targets versus their own decision-making objectives. We found that, although workers at this agency had been using the ADS continuously for nearly half a decade, the system remained a source of tension for many workers, who perceived the system's current design as a missed opportunity to effectively complement their own abilities. Overall, our findings complicate narratives from prior literature about how the AFST and similar ADS tools may fit (or fail to fit) into workers' day-to-day decision-making. These findings add to ongoing discussions in the literature,

pointing to the need for a broader re-consideration of how ADS should be designed, evaluated, and integrated into future public sector contexts [12, 182, 359].

In this section, we summarize each of our main findings, discuss how they extend or contrast with prior literature, and provide implications for the study of human-AI decision-making in child welfare and beyond. Recent work in the area of human-AI interaction has highlighted that in many real-world decision-making settings, humans and AI systems may have access to *complementary* information, opening potential for each to help overcome the other's limitations and blindspots (e.g., [11, 95, 169, 235]). We found that in the AFST context, to calibrate their reliance on the AFST score in particular instances, **workers rely heavily upon their own knowledge of qualitative, contextual details of a given referred case, which are not captured by the administrative data that the AFST model uses.** We observed that workers frequently attempted to compensate for gaps and limitations in algorithmic predictions by drawing upon rich, causal narratives they had formed about a referred case, using their unique understanding of case-specific context. These qualitative findings validate hypotheses posed in earlier work by De-Arteaga et al. (2020). Through retrospective quantitative analyses of AFST-assisted decisions, these authors found that workers were able to reliably detect and override erroneous AFST recommendations. To explain how workers were able to do this, the authors speculated that the workers may have been cross-checking algorithmic outputs against other relevant information that was available to them but not to the AFST model [95]. By contrast, much prior research on human-AI decision-making has focused on settings—often in the context of online crowdsourcing or laboratory-based studies—where the AI system has access to a *superset* of the task-relevant information that is available to the human [20, 95, 275, 416]. Our findings suggest caution in generalizing findings from such studies, as this setup may artificially advantage AI models over human decision-makers. Future research should seek to further understand how human decision-makers use complementary, model-external knowledge to calibrate their reliance on AI recommendations, as well as how future ADS tools might be designed to support them in doing so effectively. In addition, as discussed below, more research is needed to understand how human experts can be supported in *communicating* relevant contextual knowledge to an AI model, to inform algorithmic recommendations in a given case (cf. [92])—a desire expressed by workers in our study.

Despite having used the AFST day-to-day for multiple years, we found that **most workers knew very little about how the AFST works, what data it relies on, or how to work with the tool effectively.** Workers had received minimal training on how to use the AFST, and had almost no formal insight into how the AFST worked (Section 2.4.2). These findings align with recent discussions in the human-AI interaction literature, suggesting that ADS tools are often introduced into professional contexts without adequate onboarding and training for the human decision-makers who are asked to work with them day-to-day (e.g., [18, 55]). Allegheny County is outwardly transparent about the AFST's design and development process, providing much public-facing documentation on the web (e.g., [189, 436]). Indeed, this level of openness has contributed to the amount of influence and attention the AFST has received from researchers in HCI and machine learning, the popular press, and other public sector agencies around the world. However, this external transparency is not mirrored internally: workers are intentionally given limited information about the AFST model to avoid facilitating 'gaming the system' behaviors, where workers strategically manipulate data inputs to the algorithm, in order to get a desired output. Although this is a real concern in algorithm-assisted child welfare decision-making

contexts (e.g., [359]), we observed that **given minimal information about the AFST, workers improvised ways to learn more about the tool themselves, building up sophisticated yet imperfect intuitions about the AFST’s behavior.** Workers drew upon their beliefs about the AFST to calibrate their reliance on algorithmic recommendations in particular instances—including what predictive features the AFST model uses, what influence each feature has, the kinds of cases for which the model is likely to be more or less reliable, and so on. Given this influence of workers’ informal beliefs about the AFST upon their use of the tool, it is unclear whether the agency’s original goals behind limiting transparency into the model were actually achieved. Yet this lack of transparency had other consequences: workers tended to be distrustful of the tool overall, and they did not feel equipped to work with it effectively.

To learn about the AFST’s behavior in the absence of formal training, workers engaged in various forms of what Shen, DeVos et al. (2021) have called “*everyday algorithm auditing*”: the ways that users detect, understand, and interrogate machine behaviors via their day-to-day interactions with algorithmic systems [378]. Whereas these authors focused on characterizing the ways that members of large online communities come together to collectively make sense of algorithmic behavior, our findings present a case study of workers engaging in similar behaviors within a more intimate organizational setting. While readers may come away from prior quantitative research on AFST-assisted decision-making with images of workers as lone decision-makers, working in isolated cubicles, we found that workers often discussed the AFST model’s behavior with each other, to collectively understand unexpected scores. Through this collective deliberation, **workers shape each others’ beliefs about the model over time and influence each others’ decisions for specific cases** (Section 2.4.2). Workers discussed AFST scores amongst themselves frequently and informally, even though the official screening protocol assigns only one call screener and one supervisor to attend to a given case. Workers also engaged in collaborative “guessing games,” honing their abilities to predict the AFST score by guessing what the AFST score would be in particular cases, and then viewing the actual score. To learn about the influences of particular features on the AFST score, workers also frequently tinkered with the tool, for example by adjusting the inputs to the AFST model, recomputing the AFST score, and examining how the score changed. Finally, we observed that workers drew upon their knowledge of context and qualitative details of a given referred case in order to learn about the AFST’s accuracy over time—though it is important to note that the notion of “accuracy” experienced by workers was not always aligned with that of the model. We expect that workers in other public sector contexts may engage in similar strategies to build up collective understandings of an ADS’s capabilities, limitations, and overall functioning—perhaps especially in contexts where workers are *expected* to make AI-assisted decisions collaboratively, but are lacking relevant information about the AI model (e.g., [359, 475]). Prior research investigating ways to improve AI-assisted decision-making, both in the AFST context and beyond, has often focused on improving interactions between a single human and an AI system. Our findings highlight the importance of studying how *collaborative* decision-making—particularly in real-world organizational settings, in the presence of existing interpersonal and power relations among decision-makers—impacts how people rely upon and make sense of AI models.

Consistent with our findings that workers were generally skeptical and distrustful of the AFST, we observed that workers’ decisions to follow or contradict the AFST score were often guided by factors other than their “trust” in the AFST model.

Whereas prior work has often studied human reliance on ADS tools as a function of an individual's trust [20, 47, 258, 328], this conceptualization overlooks the potential impacts that organizational pressures and incentive structures may have on workers' reliance, independent of their trust in the technology itself. In the AFST context, **workers perceived organizational pressures, both from the administration and from caseworkers downstream of their decision-making, that influenced their reliance on the AFST** (Section 2.4.3). On the one hand, workers described complying with the AFST's recommendation, against their own best judgment, to avoid being accused by the administration of disagreeing with the AFST "too often" or of "not using the tool." On the other hand, workers considered caseworker load when deciding whether to screen out some ambiguous cases. Future research should seek to better understand the impacts that organizations' internal messaging, performance measures, and policies have on the ways workers rely upon ADS tools in practice. In Section 2.6, we discuss potential model-level and organization-level design implications based on these findings. In addition, based on their experiences, **workers came to perceive that they had little agency to shape the way the AFST is used within their organization, or to improve the reliability and accuracy of the AFST model itself**. As a result, workers shared that they have critical discussions about the AFST's behavior with their peers less often than they used to. Workers also believed that, overall, they are now less likely to disagree with the AFST than they once were, having acclimated to using the tool and having come to believe that their disagreement may be looked down upon by the administration. In alignment with recent calls for greater consideration of temporal factors in the study of human-AI decision-making (e.g., [250]), our findings suggest that future research should investigate how human reliance upon ADS tools may evolve over time, particularly in real-world organizational settings.

Importantly, we observed that despite having minimal formal training around the AFST, workers *were* aware of **misalignments between the AFST's predictive targets and their own decision-making objectives**, and that workers **took these misalignments into account when making AFST-assisted decisions** (Section 2.4.4). Whereas workers generally focused on ensuring children's immediate safety and considering short-term risk when making their decisions, the AFST was intentionally designed to *complement* workers' near-term focus by predicting proxies for adverse outcomes over a much longer (two-year) timespan [92, 429, 436]. The goal was to nudge workers to consider not only immediate risk and safety, but also a forecast of longer-term risk, calculated based on available administrative data. However, we found that **it was unclear to workers how exactly they were expected to take the AFST's assessments of long-term risk into account in ways that complemented their own judgment**. While recent research in HCI and machine learning has begun to explore how to support human-AI complementarity in decision-making—i.e., configurations of humans and AI systems that yield better decisions in combination than either could achieve alone [20, 95, 169, 213, 416]—our findings drive home that it is not enough to have model targets that complement workers' decision-making objectives *in theory*. Human decision-makers must also understand, at a detailed level, how an ADS tool can actually complement their abilities in practice. As Saxena et al. (2021) emphasize [359], human decision-makers must also be "bought in" to the idea that their use of a given ADS tool will improve decision-making in ways that are meaningful to them. In turn, this requires that, *even if* an AI tool is intended to complement human abilities by pushing them to consider complementary goals or objectives (cf. [171]), **there must still be value alignment between humans and the AI model at a high-level**. Yet in the AFST context, we observed fundamental

human-AI value misalignments. While some workers were open to the idea that the AFST could complement their judgment and help to mitigate some of their personal biases, other workers disagreed with the very idea of making screening decisions based on predictions of longer-term risk, perceiving this problem formulation as fundamentally misaligned with their roles and responsibilities. These observations echo findings from Saxena et al. (2021), who observed that, in agencies using simpler forms of algorithmic decision support, workers sometimes perceived conflicts between algorithmic recommendations and the ways they are actually trained to do their jobs. In addition to disagreeing with the timescale on which the AFST makes its predictions, workers were also uncomfortable with the AFST's use of proxy targets such as re-referral and placement, viewing the use of these proxies as misaligned with their personal values.

Building upon prior approaches aimed at understanding and addressing misalignments between different stakeholders' values during the design phase for ADS tools (e.g., [182, 385], future research should further explore **how diverse stakeholders (e.g., decision-makers, affected populations, data scientists, and domain experts) can be effectively engaged in shaping model-level design choices for an ADS**, such as specific choices of predictive targets or performance measures [174, 404, 474]. Future research should also explore **how best to support complementary human-AI performance in practice, in real-world organizational contexts**. This could involve, for example, investigating what other prerequisites and enabling conditions exist for human-AI complementarity, considering factors not just at the level of an AI model or individual human behavior, but also factors at the levels of groups of people or organizations. Finally, our findings suggest that when studying predictive accuracy or effective decision-making in the AFST context and other public sector decision-making contexts, it is **critical to account for potential differences in predictive targets and decision objectives of human decision-makers versus AI systems**. For instance, the misalignments observed in the AFST context complicate any argument that either the AFST or human workers are more accurate. Without an understanding of workers' objectives when making predictions and decisions, these kinds of comparisons risk evaluating workers' performance on a task that they are not actually performing.

2.6 Design Implications

Based on our findings, we provide the following design implications, intended for public-sector agencies deploying or maintaining ADS tools and for researchers exploring ways to design more effective ADS. We note that the following design implications are relevant only when there is good reason to expect benefits of having an ADS in the first place, which outweigh potential harms [25, 144, 326].

- Support workers in using their expertise to improve an ADS's performance, whether by providing direct feedback to the ADS (e.g., communicating relevant contextual knowledge to inform algorithmic recommendations, for a given case), or by allowing the ADS to learn from important patterns across multiple workers' override decisions over time, as explored in De-Arteaga et al. (2021) [92].
- Design training tools and interfaces that support workers in understanding the *boundaries* of an ADS's capabilities (e.g., where it is likely to err and what factors it *is not* able to take into account) [18, 55, 295]. For example, decision

support interfaces might be explicitly designed to support workers in exploring possible reasons for discrepancies between an ADS's statistical reasoning versus their own clinical judgment in a particular case (e.g., addressing workers' desires to have the AFST communicate back additional context explaining the score, as described in 2.4.2 and 2.4.4).

- Support open cultures for critical discussion around AI-assisted decision-making amongst workers using ADS. Providing transparency on ADS goals and algorithmic processes, as discussed above, can enable workers to engage in better informed discussions, decreasing the risk that such conversations will simply propagate misleading folk theories about an ADS. Moreover, supporting collaborative sensemaking and decision-making can encourage workers to surface insights about potential ADS limitations and improvements.
- Provide workers with balanced and contextualized feedback on their decisions. Feedback should capture the collaborative decision *process* between workers and the ADS alongside the decision *outcomes*. Both forms of feedback can be extremely valuable for learning, yet are scarce in many AI-assisted public-sector decision-making contexts (including the AFST context). Any resulting measures should be communicated to workers with explanations that account for both strengths and shortcomings of workers' judgements. Such feedback may be critical for workers to reflect upon and improve their judgment and algorithmic reliance over time.
- Relatedly, find ways to co-design measures of decision quality with the workers themselves. By designing measures of decision quality that workers find personally meaningful and important, workers may be more motivated to actually use these measures to reflect upon and improve their own decision-making. In addition, workers may be more likely to notice potential shortcomings in measures of decision quality, and may be well-equipped to notice when a given measure risks creating counterproductive incentive structures (e.g., that influence their reliance practices, as discussed in Section 2.4.3).
- Clearly communicate, and in points of disagreement, collaboratively determine with social workers how decision-making power should be distributed across workers and the ADS. Building an organizational culture in which human decision-makers feel that their expertise is relevant and valued may preface building effective human-AI partnerships. Therefore, decision-making roles between workers and ADS should be guided by an empirical understanding of workers' unique strengths (and limitations) in complementing the ADS.
- Explore methods to support diverse stakeholder involvement in shaping an ADS's model-level design decisions. For example, find methods to co-design an ADS's predictive targets with decision-makers, affected community members, data scientists, and other domain experts, while considering how to mitigate disagreements [182, 385], especially when prominent power dynamics are at play. Diverse stakeholder deliberation and discussion on model-level design decisions may help mitigate value misalignments between models and relevant stakeholders before the ADS is deployed.

- Factor in a set of work context-specific considerations when designing ADS predictive targets that aim to complement a worker's decision targets and processes. For example, beyond considering whether a human's and an AI's limitations and strengths might complement one another in a vacuum, consider the organizational factors that may influence reliance decisions in practice.
- When assessing the effectiveness of human-AI decision-making, be cautious not to adopt evaluation approaches that artificially disadvantage either the human or the AI. For example, carefully consider whether the metrics used to evaluate human-AI performance are evaluating the actual outcomes that the human versus the AI is optimizing towards, when making predictions or decisions.

2.7 Conclusion

Our findings highlight critical opportunities for future research, towards re-thinking and re-designing the interfaces, models, and organizational processes that shape the ways ADS tools are used in child welfare and other public sector decision-making contexts. Future research across HCI, machine learning, and social work should explore design opportunities at all three of these levels, both separately and in combination. In addition, future research should seek to inform the design of human-AI partnerships in child welfare by understanding and reconciling design values and goals across a broader ecosystem of relevant stakeholders, including social work administrators and agency leadership, caseworkers [359], families and other affected community members [42, 64], and ADS developers [182]. In particular, future research should explore how multiple stakeholders within this ecosystem might be meaningfully involved across various points of the design, development, deployment, use, and maintenance lifecycle for ADS, to resolve value misalignments and to better serve the needs of families and child welfare workers.

Worker-AI Interaction, Part 2

How Workers Reduce Racial Disparities in AI-Assisted Decisions

This chapter is taken from the following paper:

Hao-Fei Cheng*, Logan Stapleton*, **Anna Kawakami**, Venkatesh Sivaraman, Yanghui Di Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. “How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions.” In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*.

Abstract. Machine learning tools have been deployed in various contexts to support human decision-making, in the hope that human-algorithm collaboration can improve decision quality. However, the question of whether such collaborations reduce or exacerbate biases in decision-making remains underexplored. In this work, we conducted a mixed-methods study, analyzing child maltreatment call screen workers’ decision-making over a span of four years, and interviewing them on how they incorporate algorithmic predictions into their decision-making process. Our data analysis shows that compared to the algorithm alone, human caseworkers reduced the disparity in screen-in rate between Black and white children from 20% to 9%. Our qualitative data show that caseworkers achieve this by making holistic risk assessments and complementing the algorithm’s limitations. These results shed light on potential mechanisms for improving human-algorithm collaboration in high-risk decision-making contexts.

2.8 Introduction

Recent years have seen the deployment of AI-based tools either to augment or replace human judgments across a growing range of high-impact decision-making contexts, such as social work, criminal justice, hiring, healthcare, and education [55, 73, 148, 180, 182, 261, 323, 444, 460]. These technologies have often been adopted under the logic that they are more accurate and equitable than human decision makers [73, 209, 261]. Prior work suggests that on various predictive tasks, AI systems are more accurate than human decision makers [109, 118, 235, 423]. However, in many social decision-making contexts, such as recidivism risk assessment, AI systems have been shown to inherit human biases from historical data, and perpetuate discrimination against already vulnerable populations, e.g. [51, 52, 109, 334, 421]. Prior attempts to make these algorithms less discriminatory have largely focused on the technical design of the algorithms—a central focus of the area of *algorithmic fairness*, e.g. [110, 155]. Another possible approach to improving fairness in decision-making may be through *human-AI collaborations*, aimed at combining strengths and mitigating limitations in both AI-based and human decisions [95, 170, 280, 417, 451]. In some contexts, human-AI collaboration has demonstrated potential to improve fairness and effectiveness of decision-making, compared with human or AI decision-making alone, e.g. [20, 95, 180, 229, 322]. However, empirical results have been mixed. For example, in a real-world pretrial criminal context, human-AI collaboration was shown to exacerbate discriminatory decision-making [7, 400].

In this paper, we examine how call screen workers in Allegheny County, Pennsylvania use the AI-based Allegheny Family Screening Tool (AFST) to make decisions about which reports of child abuse or neglect (henceforth *referrals*) to investigate. Similar to De-Arteaga, Fogliato, and Chouldechova [95], we compare *automating* versus *augmenting* decision-making with the AFST. Note that the AFST was designed to “*augment the human decision whether to investigate a call*” [67] and CYF leadership assures that the tool is never used to completely automate decisions [192]. However, critics such as Eubanks [119] and the Child Protection Reform [69] worry that the AFST, and other tools like it, may someday be used to automate decisions, for example, as an austerity measure. In this paper, we investigate the effect that such a hypothetical automated decision-making policy would have on racial disparities in child welfare call screening. We then compare this policy to the current standard decision-making process, where workers make decisions with the assistance of AFST. Since child welfare workers themselves are known to make racially disparate decisions [100, 101, 108, 152, 232], it is unclear whether adding them back “in the loop” will do any good in this regard. The central question of this paper is thus: **when people work with algorithms in a child welfare context that is known to have racial disparities, will they serve to mitigate or exacerbate disparities?** The answer to this question can inform the responsible design and use of AI tools in the child welfare context, as well as other high-stakes social decision-making contexts.

The primary racial disparity measure we use in this paper is the difference in the screen-in rates across Black and white children. Where prior work has emphasized the impact that the AFST has had on caseworkers’ workloads [140], we think it is also important to focus on the impact that the AFST has on families in Allegheny County. Call screening is most often the first point of contact with the child welfare system, where an agency decides whether to intervene into the life of a family by investigating them. Being screened in may lead to more child welfare involvement into a family’s life. AFST documentation states that “*screening in and a child protection investigation has some potentially deleterious effects on families. If screening in, however,*

is a prerequisite to being offered higher quality services or being prioritized for a slot in a desired program, one can argue the benefits of an investigation” [431]. Higher screen-in rates indicate higher levels of state intervention into families’ lives, starting with investigation. Racial disparities in screening rates indicate uneven application of interventions or investigations of Black and white families and potentially uneven distribution of the potential harms and benefits of them. Disparities in screening rates may be unjustified if they occur because of unwarranted intervention or lack of intervention, e.g. investigating a family when their child is not at imminent risk of abuse or neglect. As Dorothy Roberts [105] suggests, “[t]he disproportionate number of Black children under state supervision results from discriminatory decision-making within the system as well as racist institutions in the broader society.” However, as we will discuss below, disparities may be justified: higher need or higher risk of abuse or neglect among children in one group of people could warrant a higher screen-in rate.

In summary:

- Through quantitative analysis based on the two years of data immediately following the introduction of the AFST—from August 2016 to July 2018—we evaluated racial disparities in AFST-only and worker-AFST screening decisions. Our results show that worker-AFST decision-making served to reduce the disparity in screen-in rate between Black and white children compared to algorithm-only decisions.
- We conducted a contextual inquiry by observing how call screen workers use the AFST to help them make decisions, and we interviewed workers about their experiences working with the algorithm. Through qualitative analysis, we find that by assessing referrals holistically using all of the information available to them and by adjusting for the algorithm’s limitations, call screen workers disagreed with the AFST in ways that serve, in aggregate, to reduced the impact of racial disparities in the algorithm. Our findings suggest that the AFST did not supplant call screening discretion and decision processes and that workers were not blindly following the AFST, consistent with the statement from Human Services [192] responding to Eubanks [119].
- We also analyzed the accuracy of AFST-only and worker-AFST decisions. Although the AFST is better than workers at predicting the outcomes that it is trained to predict, our qualitative findings indicate that workers make screening decisions to optimize for fundamentally different outcomes than the AFST.
- Finally, we provide design implications for potential ways to improve the collaboration between call screen workers and the AFST in improving the decision-making process.

The AFST is just one algorithmic system used in child welfare; there are many similar systems used in child welfare agencies across the U.S. [354, 361]. We anticipate that several of our findings may generalize to other public sector, algorithm-assisted decision-making contexts. However, agencies are often reticent about their internal policies, decision-making, and even public information [2, 358], making it challenging to conduct similar analyses across other contexts. We thank the Allegheny County Office of Children, Youth and Families (CYF) for their continued interest in working with external researchers, and for their transparency in providing us data and access to their facilities. We also thank the workers in the Intake Department for taking the time to speak with us, and for sharing their insights. We

hope that more agencies will follow Allegheny County’s lead in opening opportunities for public and research accountability.

Finally, the data used throughout this paper contained information on 56,843 children who were referred to CYF. We acknowledge all 56,843 of these children and their families, on whom this data was collected and for whom these records reflect a potentially consequential interaction with CYF.

2.9 Related Work

2.9.1 Algorithm-in-the-loop decision-making

“Algorithm-in-the-loop” decision-making is commonly characterized by first having an algorithm-produced prediction or classification, with a human making the final decision after considering an algorithmic recommendation [148, 401]. Algorithm-in-the-loop decision-making has been observed in multiple high-stakes scenarios, including pretrial bail decisions [7], recidivism predictions [109], predictive policing [324], and diagnosing patients in clinical settings [293].

As algorithm-in-the-loop decision-making becomes increasingly common in practice, recent research has started to look at how humans work with algorithms when making decisions and at the relative contributions of humans versus algorithms to overall performance. Many studies have focused on prediction accuracy, finding that on many tasks, algorithms can outperform humans in terms of prediction accuracy [45, 251, 252, 471]. However, recognizing that human experts and algorithms may have complementary strengths and limitations, a line of research has sought to understand how to combine the capabilities of each [95, 170, 280, 417, 451]. Some studies have demonstrated that combinations of human and algorithmic judgment can improve prediction and/or decision-making (e.g., [20, 95, 180, 229, 252, 322]). Yet empirical results in this space have been varied so far. In other studies, human-algorithm decision-making has failed to improve or has even *harmed* performance, compared with either human or algorithmic decisions alone (e.g., [7, 146, 329, 400, 471]).

Beyond accuracy, other metrics have been used to evaluate between decisions made by humans, AI, and human-AI combined [7, 123, 148, 226, 273, 329]. Much of this work focuses on minimizing the error of the decisions compared to the ground truth. Most relevant to our work, Albright [7] compares racial disparities in human-only and human-AI decisions in the context of pretrial bail hearings. Here, judges are presented with a risk score and recommendation from a risk assessment algorithm (similar to the AFST) and then must decide whether to give a person who is charged with a crime bail (and keep them in jail until they can pay) or let them go free without paying bail until their trial date. Albright [7] suggested that judges disagreed with the algorithmic recommendations for certain types of defendants such that the judge-algorithm bail decisions were more racially disparate—giving Black people bail rather than letting them free without bail at a higher rate than white people—than both the past judge-only decisions (before the algorithm was implemented) and the algorithm-only recommendations in the same time period. In our work, we present empirical results in the opposite direction: in the context of Child Welfare call screening, we found that human-algorithm decisions were less racially disparate than algorithm-only decisions would have been, and somewhat less racially disparate than past human-only decisions.

2.9.2 Risk Assessment Tools in Child Welfare

For an overview of predictive algorithms used in child welfare, see [361] and [354]. We provide a brief history here. For decades, child welfare agencies across the U.S. and abroad have been using risk assessment instruments (RAIs) to assist child social workers in making decisions, such as whether or not to investigate a family or whether to remove a child from their family. Most RAIs have been checklists that workers fill out in order to estimate the risk of child maltreatment. For example, see the Structured Decision Making (SDM) tools used in the California Child Protective Services system and a number of other locales [87]. For a case study of other kinds of RAIs and algorithms used in child welfare, see [360]. However, newer RAIs include automated tools, commonly called *predictive analytics* or *data-driven predictive tools*, which use statistical modeling and machine learning to estimate risk based on historical administrative data. Earlier iterations of these were developed by private companies, such as Eckerd Connects [78], MindShare Technology [419], or SAS [195]. Due to high error rates and their “black box” nature, the Los Angeles County and Illinois child welfare systems dropped private algorithms after brief trials [68, 200, 304]. As such, these private tools have fallen out of favor, though some were still in use in other locations at the time of publication. Other data-driven predictive tools have been or are being developed through public-academic partnerships — such as the Allegheny Family Screening Tool (AFST) in Allegheny County, Pennsylvania [73, 431] or other tools being implemented by the same designers in Douglas County, Colorado [430] and Los Angeles County [343].¹ These publicly-developed algorithms have proved more resilient, with the AFST being the longest-lasting and most prominent predictive tool in use today. Proponents of these newer automated risk assessment tools, such as the AFST, argue that they make more accurate decisions than both child social workers and standard checklist-based RAIs; and that they make more consistent, objective, equitable decisions [73, 88, 193, 393]. Critics argue that these automated tools are still too inaccurate, that they do not predict true child abuse or neglect, and that they still make biased decisions because they are trained on biased data [70, 73, 119].

In order to evaluate the AFST based on these opposing concerns, the Allegheny County CYF commissioned Goldhaber-Fiebert and Prince [140] to conduct an Impact Evaluation of the AFST. Among the results, the most relevant to this paper were the following: 1) the AFST “*increased accuracy for children screened-in for investigation and may have slightly decreased accuracy for children screened-out;*” 2) the AFST did not decrease the screen-in rate overall; and 3) the AFST reduced racial disparities in terms of the screen-in rate, but possibly worsened them in terms of accuracy [140]. These findings called into question whether the AFST improves either accuracy or equity. These results were based on an analysis of two years of data immediately following the introduction of the AFST—from August 2016 to July 2018. Our quantitative findings are based on the same data as Goldhaber-Fiebert and Prince [140], which we preprocessed and analyzed to match their work, as well. Rather than reiterate their findings, we use this data to evaluate the effects that automating screening decisions would have on racial disparities. We also expand upon Goldhaber-Fiebert and Prince [140]’s findings with a new mixed-methods approach.

¹It is notable that Los Angeles County, the largest child welfare department in the U.S., dropped a private model earlier, but is now implementing a new model with the same team who developed the AFST.

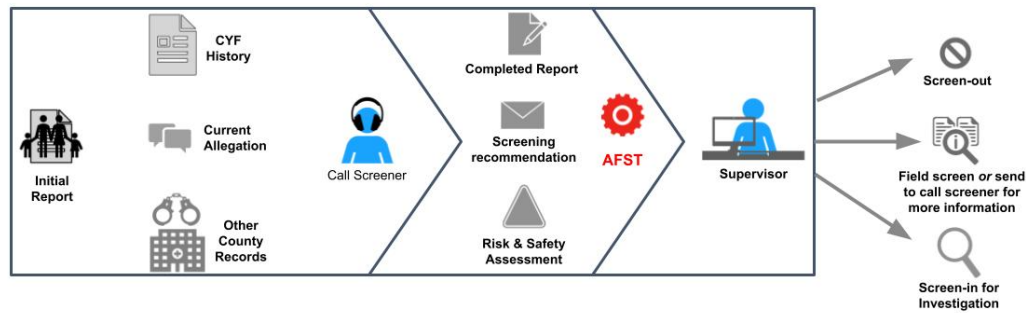


FIGURE 2.2: The current AFST-assisted call screening process. Call screen caseworkers make screening recommendations and supervisors make the final screening decisions, both with the AFST’s risk score and recommendation.

2.10 Methods

2.10.1 Study Context

In this paper, we studied a high-stakes scenario of child maltreatment referral screening decisions. The Allegheny County Office of Children, Youth, and Families (CYF) has been using an algorithmic tool, the Allegheny Family Screening Tool (AFST), to assist with child maltreatment call screening process since August 2016. The AFST is a machine learning-based predictive risk modeling tool that analyzes county data to predict outcomes related to child abuse or neglect. The original version of the AFST used demographic data related to the alleged victims, caretakers, alleged perpetrators, prior child welfare history, criminal history, public behavioral health history, and use of public assistance [431].² Prior to the introduction of the AFST, caseworkers and supervisors made all referral decisions without any algorithmic aids. Since its deployment, caseworkers and supervisors have been presented with an AFST risk score to assist with their call screening decisions for all referrals which were not automatically screened in or out (henceforth *referrals* or *discretionary referrals*).³ Call screen workers still make the final decisions—they have the option to either agree with the algorithm recommendation, or to disagree and go with their own decisions.

We adopted a mixed-methods approach to investigate how CYF call screen workers work with the AFST, and how the resulting human-AI decisions affect disparities in decision outcomes. We analyzed historical data on call screening decisions prior to and after the deployment of the AFST algorithm. We also conducted contextual interviews with caseworkers and supervisors to support interpretation of findings from these quantitative analyses.

Use of the AFST in call screening

The AFST Version 1 (used from August 2016 to November 2018) was made up of two models—the *re-referral* model was trained to predict whether a child would be reported again within two years of being screened out; the *placement* model was trained to predict whether a child would be removed from their home and placed in foster care within two years of being screened in. Each model produces a risk score

²This was true for the first deployed version of the tool. The second version (in use from November 2018 until the time of publication) stopped using public assistance as a predictive feature and started using birth records [432].

³We provide more detail about which referrals are discretionary or not in Section 2.10.2.

ranging from 1 (lowest risk) to 20 (highest risk) associated with the likelihood of the corresponding predicted outcome (re-referral or placement) for each child in the referral. The score is categorized visually into 3 bins: Low risk (score 1-9), Moderate risk (score 10-14), and High risk (score 15-20). The caseworker and supervisor sees a single risk score (see Figure 2.3) associated with the referral. The presented score is the higher of the scores across the two models. If the referral involves multiple children, the score of the child with the highest score in the referral would be shown.⁴

Workers used the AFST as follows: each caseworker first gathered information about the referral, made assessments of risk and safety,⁵ made a screening recommendation for the referral, then they ran the AFST to generate a risk score and passed the report to a supervisor for review. For most referrals, the AFST score served as a recommendation, and workers had the authority to either agree or disagree with that recommendation when making the final screening decision. In Appendix ??, we suggest that CYF workers consider High risk referrals (15-20) to be recommended screen-in, Low risk (1-9) recommended screen-out, and Moderate risk (10-14) sans recommendation. However, 29.3% of children in discretionary referrals from August 2016 to May 2018 had a placement model score of 18 or above and were flagged as *mandatory screen-in* [431]. For these referrals, workers were shown the AFST interface on the right side of Figure 2.3. These referrals were “required to be screened in,” but supervisors were able to “override this requirement at their discretion” provided that they “documented and reviewed” their reasons for overriding this requirement [431]. Supervisors overrode 21.0% of these children in *mandatory screen-in* referrals.

2.10.2 Data analysis

We acquired data on all child maltreatment referrals at Allegheny County CYF from January 2015 to July 2018. The data contains all alleged child victims that were referred to the county child protective service within the time span. The data contain referrals from both before the deployment of AFST (January 2015 - July 2016), and after the deployment of the AFST V1 model (August 2016 - July 2018). We only considered data after January 2015 due to a change in the child maltreatment policy in the state, introduced on January 1st 2015. We also only consider general protective services (GPS) referral in the time span (since referrals under the other category, child protective services (CPS), are all required to be screened in).

The data contains 37,065 records of referrals before the AFST deployment, and 59,039 records after the AFST deployment. Each record contains information on the primary alleged victim(s) of the referral, with features that are used by call workers and algorithms in making call screening decisions. Each record also includes the AFST risk score generated for the alleged victim(s) in each referral, and the final call screening decisions made by the call workers for that referral. For our analysis, we look at the outcome at an individual level. We further narrowed

⁴For example, if there are two children in the referral: the first child with a re-referral model score of 1 and a placement model score of 15, the second child with a re-referral score of 10 and a placement score of 3; a score of 15 (and a “High risk” message along with Figure 2.3) would be shown to the call screen workers.

⁵This is the risk of future child maltreatment and immediate safety. A U.S. government source defines a safety assessment as gathering information to “determine the degree to which a child or youth is likely to suffer maltreatment in the immediate future” and a risk assessment as collecting information “to determine the degree to which key factors are present in a family situation that increase the likelihood of future maltreatment to a child or adolescent” [134].

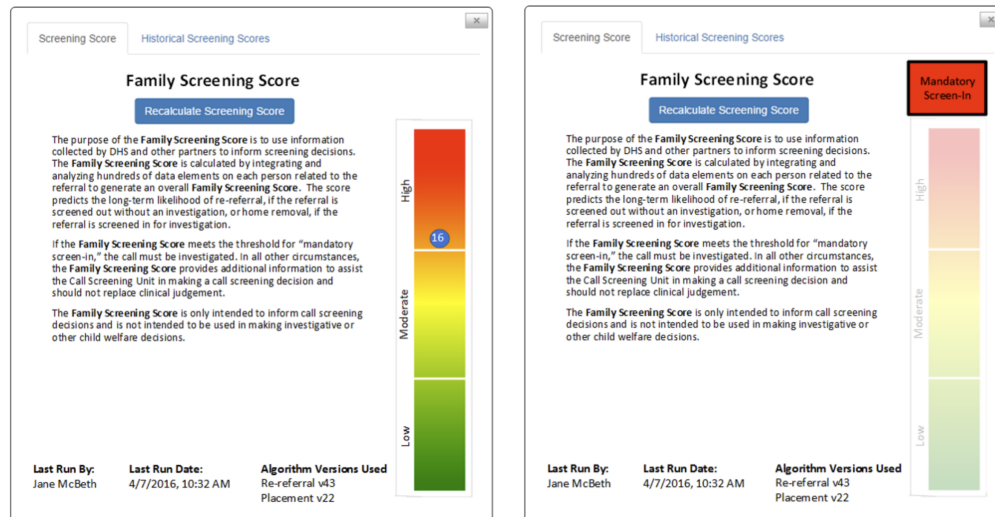


FIGURE 2.3: The interface of the AFST which is used by the child welfare workers to make screening decisions. The left figure shows the score the for a normal referral. The right figure shows a referral with an exceptional high risk that triggers the mandatory screen-in policy.

down the referrals to those for which call screen workers had some discretion over, which yield us 31,025 referrals before the AFST deployment, and 51,750 records after AFST deployment. Some children were referred multiple times to CYF, generating multiple records. After pre-processing, our data contained referral information on 23,230 unique children pre-deployment and 33,613 children post-deployment, totalling 56,843 children on whom the data used in this paper was collected.

As the AFST was put into deployment in August 2016, all call screening decisions made after August 1st 2016 were collaboratively made by call screen workers with the assistance of AFST. For the data before August 2016, the call screening decisions were made solely by the call workers without any algorithm aid. However, AFST scores were calculated retrospectively for the referrals. Thus all referrals from January 2015 to July 2018 have AFST scores.

2.10.3 Contextual Inquiry and Interviews

To support interpretation of our quantitative findings, we also conducted contextual interviews with call screen caseworkers and supervisors. A group of researchers visited the Intake Department of Allegheny County CYF in July 2021. The visit consisted of two parts: 1) **contextual inquiries**, where the researchers observed how call screen workers worked with the AFST when making screening decisions; and 2) **semi-structured interviews** where researchers were able to ask more in-depth questions. We observed and interviewed 13 participants in total: 9 call screen caseworkers and 4 supervisors over 2 separate visits in a span of two weeks. All participants worked full time as call screen caseworkers or supervisors. We checked with workers before and during our visits to make sure that we did not burden them too much while they were busy with work. To prevent workers from being identified in the workplace, all responses in the paper are anonymous and we report only minimal demographic information. Participants included, but may not have been limited to, white workers, Black workers, women, and men. At the request of the office, we did

not provide monetary compensation to the participants. This study was approved by the Institutional Review Board of Carnegie Mellon University.

Contextual Inquiry

After obtaining consent from participants, we observed call screen caseworkers and supervisors in their normal workflow, and participants were encouraged to think aloud as they performed their tasks, to make more of their thinking and reasoning visible. See Figure 2.2 for a visual diagram of the overall workflow. For caseworkers, this included taking phone calls from reporting sources, gathering information for reports, running the AFST to produce a risk score and recommendation, and making the call screening recommendation. For supervisors, this included reviewing the reports made by caseworkers, correcting information in reports (if need be), requesting field screening to gather missing information about a referral (if need be), making the final screening decisions, and overriding the mandatory screen-in referrals that the algorithm enforces.⁶ Each contextual inquiry session took about three hours. Researchers took notes on the actions and thought processes of the participants, while asking brief follow-up questions as needed. Due to the sensitive nature of the work, we neither audio recorded the contextual inquiry nor took notes on any personally identifiable information.

Semi-structured Interviews

After the contextual inquiry, we invited each caseworker and supervisor for a semi-structured interview. The goal of the interview was to understand how participants incorporated the AFST in their decision-making process, and in particular, to gain further insight into possible mechanisms underlying our quantitative findings. At the beginning of each interview, we discussed participants' background and experience in child welfare. We also asked any follow-up questions that arose from our observations during the contextual inquiry. We then asked participants how they worked with the AFST to ensure fairness in screening decisions. We shared statistics about racial disparities in call screening, comparing the AFST recommendations to actual decisions from 2016 to 2018, and asked workers' thoughts on these numbers. Lastly, we discussed potential improvements to the design and use of the AFST.

Qualitative Analysis

We first transcribed all interview recordings into text, and used thematic analysis [40] to analyze our data, a constructivist approach inspired by grounded theory [60]. We combined the data collected from both the contextual inquiry and interviews, which contains interview transcripts and field notes. The authors collaboratively conducted open coding on the data, which generated over 1500 open codes. The authors then conducted an iterative affinity mapping process to the open codes, performing constant comparisons and iteratively clustering related codes. In the end, the authors refined the themes that emerged from the affinity mapping.

⁶The AFST automatically flags referrals with the highest risk as "mandatory screen-in." Only supervisors have the authority to override the AFST's decisions for these referrals [AFSTfaq].

2.10.4 Positionality

We authors acknowledge that our positionality shapes our approaches to research, as well as how we interpret and present our findings. Given that the subject of research involves how Black and white families have been treated by the child welfare system in Allegheny County, Pennsylvania, we think it is especially important to acknowledge our racial/ethnic backgrounds, where we live, and our relationships to child welfare. The two lead authors are Asian and white, respectively. Neither have children nor have been investigated by a child welfare agency, nor were they adopted nor involved in the foster care system as children. All but two authors live in Allegheny County; the other two live in Minneapolis and California. To our knowledge, none of us have been significantly personally impacted by child welfare. Throughout this work, we collaborated with Allegheny County CYF in order to gain access to data and to talk with workers, but the analysis and writing of this work were conducted independently.

2.11 Decision-making paradigms and terminology

In this section, we define the terminology used throughout the paper.

2.11.1 Decision-making paradigms

In this paper, we compare hypothetical screening decisions made by the AFST alone versus actual decisions made by child welfare workers when using the AFST. We define these two decision-making paradigms as follows:

1. **AFST-only decisions:** The hypothetical screening decisions that the AFST algorithm would make if it were the only decision-maker (without workers). For clarity, we suppose that the AFST would screen in any High risk referrals (with a score of 15 to 20) and screen out any Moderate and Low risk referrals (scores 1 to 14). We choose a threshold of 15 for our AFST-only policy because it is the threshold between Medium and High risk AFST labels presented to the call screen workers, and because its hypothetical screen-in rate would be close to the actual screen-in rate from 2016 to 2018. This split between High and Low risk referrals follows official AFST documentation, which discusses disparities: “up until the end of 2017, 47% of black children received a ‘high’-range score (15–20), compared to 39% of white children. Conversely, 29.6% of white children have received a ‘low’-range score (1–9), compared to 10% of black children” [AFSTfaq]. See Appendix A.4 for evidence that call screening workers see High risk labels as screen-in recommendations and Low risk as screen-out. However, we also conducted sensitivity analyses by replicating our empirical results across alternative thresholds. For example, Figure A.1a shows similar screen-in rate disparities across thresholds from 10 to 20.
2. **AFST-assisted worker decisions:** The actual call screening decisions made by child welfare workers from 2016 to 2018, assisted by the AFST (see Section 2.10.1 for a detailed description of the decision-making process). For brevity, we refer to these decisions as **worker-AFST decisions** throughout the rest of the paper.

2.11.2 Definitions of Black and white

In the data, each child had one or more of the CYF race labels: “Black,” “white,” “Hispanic,” “Asian,” “Native American,” “other,” or “unknown.” For our quantitative analyses in this paper, we consider a child Black if they were assigned the CYF race label “Black” alone or “Black” with any other CYF race label. We consider a child white only if they were “white” alone —i.e. if the child had the CYF labels “white” plus any other race, they were not considered white in our quantitative analyses. We consider a child with the CYF labels “Black” and “white” as Black. This follows the same racial classification as the official AFST Impact Evaluation [140].

2.11.3 Evaluation Metrics

Our primary research questions to evaluate are: 1) how worker-AFST decisions affected racial disparities in call screening, and 2) whether changes in disparities affected the decision accuracy. We adopt the following evaluation metrics:

- **Racial disparity**

The primary disparity measure used in this paper is the difference in the screen-in rate between Black children versus white children. Differences in the screen-in rates between different racial groups correspond to one of the simplest and most popular algorithmic fairness notions - statistical parity, e.g. [79, 110, 122]. Specifically, a classifier satisfies statistical parity if the subjects in the protected and unprotected groups have equal probability of being assigned to the positive predicted class. While statistical parity serves as a starting place for our analyses, to assess the robustness of our results we also evaluate the disparities in other metrics, including accuracy, precision, true positive rate and false positive rate.

- **Accuracy**

We measure accuracy by the percentage of decisions that align with a proxy ground truth: for screen-in decisions, we measure the percentage of children that are either removed from home within 2 years or re-referred again within 2 months of the referral; for screen-out decisions, we measure the percentage of children who are neither removed from home within 2 years nor re-referred again within 2 months of a referral.⁷ Our definition of accuracy differs slightly from prior work evaluating the AFST, for which a screen-in decision is accurate only if the child is later placed in foster care within 2 years and a screen-out decision is accurate only if the child is not re-referred within 2 years [140, 431]. We adopt the former definition of accuracy so that our hypothetical AFST-only accuracy is a decent estimate of what the accuracy of screening decisions would have been had they actually been automated by the AFST-only policy.

Our measurement of AFST-only accuracy is an imperfect estimate, due to its counterfactual predictions [82]. For any referrals where the AFST-only decision differs from the actual screening decision, ideally we would want to measure the AFST-only accuracy in terms of counterfactual outcomes —e.g. whether a child would have been placed in foster care or re-referred had they been screened in (when in reality they were screened out). Since we do not know these counterfactual outcomes, we evaluate AFST-only accuracy based

⁷We chose to evaluate re-referral within 2 months for consistency with the official AFST Impact Evaluation [140]. It is worth noting, however, that the AFST (V1) predicted re-referral within 2 years [431].

on the actual predictive outcomes instead. However, because the screening decisions affect the predictive outcomes, the actual outcomes may have different probabilities than the counterfactual ones. For example, if the AFST-only decision is screen in, its accuracy will be judged based on whether the child is re-referred or placed; but, if they are actually screened out, we assume (but do not know) that the child is less likely to be placed and more likely to be re-referred than if they were actually screened in. This assumption may distort our measurement of accuracy for the hypothetical AFST-only decisions. As Coston et al. [82] note, this is a limitation endemic to risk assessments where the predictions affect the predictive targets.

- **Other metrics**

We also evaluated disparities and prediction performance for a few additional metrics. We defer these results and the necessary terminology to the Appendix.

2.12 How AFST-assisted worker decisions affected screen-in rate disparities

2.12.1 Black-white disparities between AFST-only decisions and workers' final screening decisions

In this section, we look at the disparity in screen-in rates among Black and white children who were referred to CYF for AFST-only and worker-AFST screening policies. In Figure 2.4, we see the difference in screen-in rates between white and Black children under these two screening policies for all children reported to CYF from August 2016 to July 2018 who were not immediately screened in or out.⁸ The screen-in rates in Figure 2.4 are calculated over the number of discretionary referrals within the race listed, e.g. the 71% AFST-only screen-in rate for reports with Black children means that 71% of reports with Black children would have been screened in from 2016 to 2018 following the AFST-only screening policy. The *Black-white screen-in rate disparity* refers to the difference between the screen-in rate for Black children and the screen-in rate for white children under a given policy for this time period. For example, the AFST-only Black-white screen-in rate disparity from 2016 to 2018 was 20%, because 71% of all referrals with Black children would have been screened in and 51% of all referrals with white children would have been screened under the AFST-only screening policy.⁹ We calculate that the worker-AFST screen-in rate disparity for the same time period was 9%, since 61.8% of Black children and 52.8% of white children were screened-in. **These results suggest that the Black-white screen-in rate disparity under the worker-AFST screening policy was less than half than that of the AFST-only policy from August 2016 to May 2018, 11% lower to be exact.** Table 2.2 shows the total number of referrals with Black and white children which would have been screened in and out under the AFST-only policy. The AFST-only decisions would have screened in over 7500 more Black than white children for their families to be investigated by CYF. In actuality, workers only screened in 4713 more Black than white children. For reference, we calculate that from January 2015 to July

⁸By Pennsylvania law, all CPS referrals are automatically screened in. Some other referrals were marked as automatically screened in or out. All of these referrals were excluded from our analysis. All numbers in this section reflect only *discretionary referrals* for which the AFST could have had some influence on.

⁹Note that this 20% difference is in percentage points, i.e. 71% minus 51%.

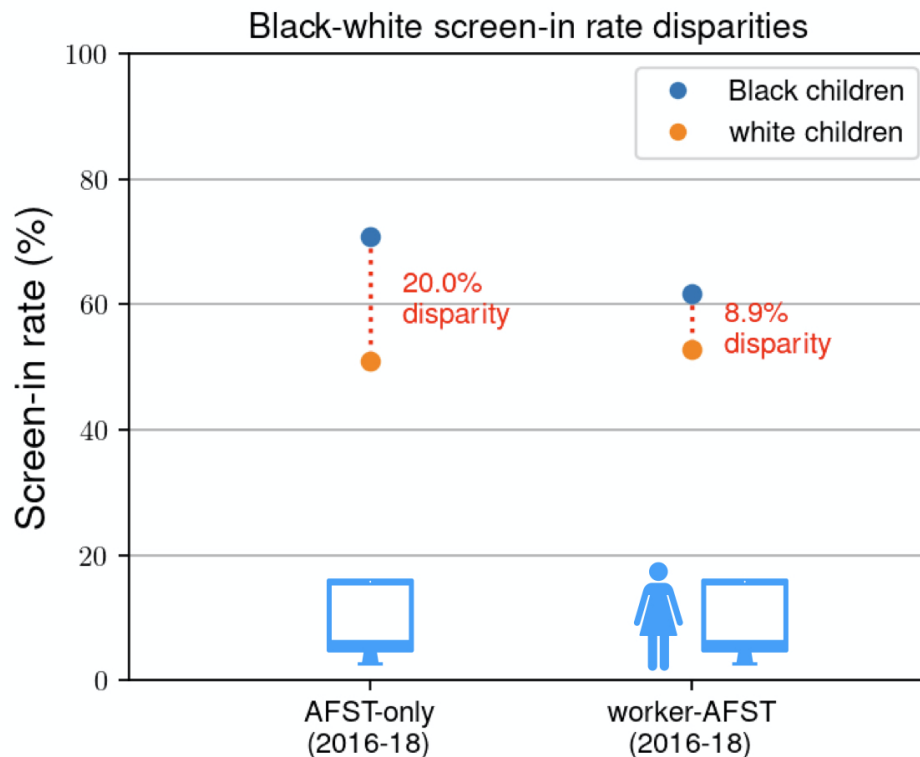


FIGURE 2.4: Black-white screen-in rate disparities for AFST-only and worker-AFST decision

	AFST-only		Worker-AFST	
	Black	white	Black	white
Screen-in	18536	11013	16133	11420
Screen-out	7587	10610	9990	10203
Total	26123	21623	26123	21623

TABLE 2.2: Total number of children that the AFST-only would have hypothetically screened in and out, compared to the actual decisions made by worker-AFST.

2016, before the AFST was implemented, workers screened in 52.5% of Black children and 41.2% of white children in discretionary referrals. This was a Black-white screen-in rate disparity of 11.3%.

We also examined Black-white screen-in rate disparities using other decision thresholds for the hypothetical AFST-only policy, as well as additional disparity metrics. We find that worker-AFST decisions were less disparate than AFST-only decisions, regardless of which threshold is chosen. It should be noted that the threshold of 15 that we default to throughout the paper has the second-to-highest disparity of any threshold from score 10 to 20 (including “mandatory screen-in”). We default to threshold 15 not to overstate our results, but because we argue that it would be the most reasonable threshold to choose from, given the design of the AFST which splits referrals into High and Medium at score 15, and given that a score of 15 produces a reasonable screen-in rate, whereas other thresholds do not. We observe similar patterns for the disparities in precision rates, true positive rates and false positive rates—worker-AFST decisions exhibit less racial disparity in these metrics, as compared to AFST-only decisions for all reasonable threshold levels. See Appendix A.3

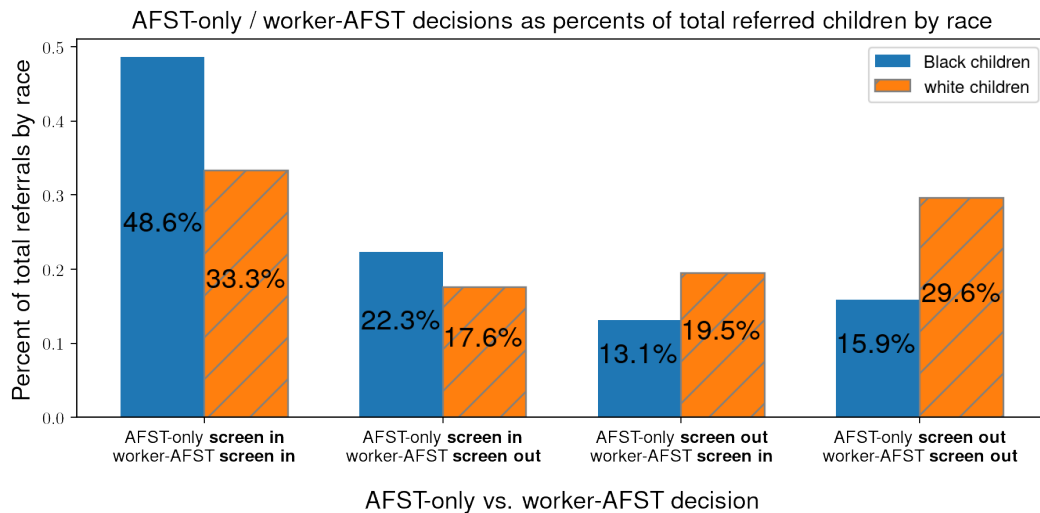


FIGURE 2.5: Proportions of Black and white children for which the AFST-only decisions and worker-AFST decisions agreed and differed. Note that the percentages here are over the total number of referrals per race, e.g. the 48.6% in the leftmost bar indicates that 48.6% of all Black children referred to CYF would have been screened in by AFST-only and was screened in by workers aided by the AFST (worker-AFST). Notice that the sum of the two leftmost bars (for a single race) equals the AFST-only screen-in rate —e.g. $48.6\% + 22.3\% = 71\%$ for Black children,— whereas the sum of the leftmost and the right middle bars equals the actual worker-AFST screen-in rate —e.g. $48.6\% + 15.9\% = 61.7\%$.

for our complete analysis.

How worker and AFST disagreement reduced the Black-white screen-in rate disparity. Figure 2.5 shows the percentage of children which would have been screened in and out under both the AFST-only and worker-AFST policies, broken down by race. Recall that the Black-white screen-in rate disparity under the AFST-only screening policy was 20%. The simplest way that workers could have reduced this disparity would have been to screen in more white families and screen out more Black families than the AFST-only policy would have. Overall, this is what we observed: however, workers did not disagree with the AFST-only policy exclusively in ways that would have lessened this disparity. If workers were heavily guided by the AFST, but intentionally tried to reduce its screen-in disparity, we might expect that workers would not screen in Black families that the AFST-only would have screened out, and that workers would not screen out white families that the AFST-only would have screened in. However, we see in Figure 2.5 that this is not the case: workers screened out 15.9% of Black children who would have been screened out under the AFST-only policy and screened out 17.6% of white children who would have been screened in under the AFST-only policy. Because we see disagreement across the board, it is likely that workers are using their best judgment across the board, and they are not simply following the AFST's recommendations. It is also likely that workers are not making screening decisions explicitly in order to reduce racial disparities. These interpretations align with our qualitative findings in Section 2.12.2.

What ultimately led to the Black-white screen-in rate disparity being lower under the worker-AFST policy than under the AFST-only policy is shown in Figure 2.5. **For children that the AFST-only policy would have screened in, call screen workers screened out more Black children than with white children (22.3% vs 17.6%). For**

referrals that the AFST-only policy would have screened out, call screen workers screened out less Black children than with white children (15.9% vs 29.6%).

2.12.2 Qualitative: How workers achieved lower Black-white screen-in rate disparity?

When reviewing a referral, workers can see the races of everyone involved. In theory, one (naive) way workers could have reduced the screen-in rate racial disparity is by simply looking at the race of the children in the referral, screening in more Black children and screening out more when white children, regardless of the other factors involved in the referral. Based on both our quantitative results (Section 2.12.1) and our qualitative findings, it is clear that this was not the case: workers were making decisions in a more sophisticated way. Furthermore, four call screen caseworkers and three supervisors explicitly said they did not make screening decisions based on the race of the family.¹⁰ Two caseworkers and three supervisors said they did not look at race at all. For example, one supervisor said, *“I have no idea what races people are.”* Based on contextual inquiries and interviews with CYF call screen workers, we hypothesize that the decrease in Black-white screening disparity from 2016 to 2018 occurred because of the following reasons:

1. Workers made screening decisions based on holistic, contextual assessments of risk and safety.
2. Workers adjusted for limitations that they perceived in the AFST when making decisions on a case-by-case basis. For some workers, this adjustment was a conscious adjustment to try to reduce racial disparities. For others, it was unintentional. However, the effect of reducing disparity in aggregate was the same.
3. Workers made collaborative screening decisions about some reports they were uncertain about.

Holistic decision-making

Workers made holistic, contextual assessments of risk and safety in order to determine whether to screen referrals in or out.

Interview findings: Four caseworkers and one supervisor said they did not consider the AFST’s recommendation as a baseline to guide their final screening decision. Rather, workers said the AFST provided additional information to consider (two supervisors and three caseworkers). For example, one supervisor said, *“[the AFST is] a good tool to have for some extra information, in terms of risk.”* Another supervisor said workers *“take in consideration what the computer [i.e. the AFST] is saying, but,... we’re not making a decision based on what the computer says. If the computer says, ‘This score is a 16,’ per the computer we have to assign that family, and we’re not doing that, we’re using real... information to make the decision and not numbers to make the decision.”* This supervisor also said they screen reports based on *“what risk factors are going to impact the children, because it’s all about child safety.”*

For call screen workers, *safety* refers to present danger or well-being of the children in a report; *risk* refers to the chance that the children will be harmed or neglected

¹⁰Note, however, that five caseworkers stated that they did consider the race of the family in order to account for racial biases in reporting. We explain this further below.

in the future. One supervisor said, *“safety is more immediate. And risk is even more long-term.”* For example, one caseworker said living in a dirty home is a risk, but not an impending danger. Workers conduct *holistic, contextual assessments* when they consider any piece of information in a report to be relevant only when it’s relevant to risk and safety in the context of all other information in the report (cf. [11]).

Contextual Inquiry Observations: For example, we observed a caseworker review a referral where a single mother was reported by a friend for allegedly using drugs and generally neglecting her four children (ages five to seventeen). The report also included pressing concerns about some of the children’s dental health. The family was reported a month prior because one of the children was truant. The AFST score was 11 (moderate, on the side of screening out), which this caseworker considered to be a low score. However, the caseworker recommended screening in the referral, because of the combination of the drug allegation, the presence of some young children, the past truancy case, and the dental health concerns. The caseworker said they would have screened out the referral if it had included only one or two of these risk factors without the others.

Overall: Because workers considered the AFST risk score in the context of all the other information in a referral, we often observed that they considered other information to be more relevant to risk and safety than the AFST score, and made screening decisions despite the score. We suspect that this holistic decision-making contributed to the baseline level of disagreement across both race and AFST-only screening decisions in Figure 2.5. One supervisor expressed that they thought this kind of holistic, contextual decision-making led to less racially disparate screening decisions. However, this supervisor said they did so unintentionally and that the 11% reduction in screen-in rate Black-white disparity from 2016 to 2018 was *“not the intention, just the outcome.”*

Adjusting for limitations of the AFST

Next, workers said they adjusted for what they perceived to be limitations of the AFST: they disregarded the AFST’s recommendations when they thought it over- or under-scored referrals because it was unable to properly take all referral-relevant information into account. Instead, workers relied on factors that they believed were more relevant in a given referral to make a decision, as discussed below.

(1) Workers thought the AFST was over- or under-scoring certain referrals because it did not take the allegation or other current referral information into account properly.

Interview findings: For example, one caseworker said that the AFST does not consider the allegation in the report when determining the risk score, *“even if it says, ‘Dad killed Mom in front of the kids.’ You know what I’m saying? Like, something crazy.”* One caseworker gave an example where the AFST over-scored a referral because it did not properly take into account the results of a recent investigation: *“We just investigated and we found that the parents are providing fine... and we just closed it. Then some anonymous person reports the same thing. Then... [the AFST says] it’s a high risk again, and we already just previously addressed it.”*

Contextual Inquiry Observations: For example, we observed a caseworker review a referral where a mother was reported for allegedly not giving prescribed mental health medicine to her daughter. The caseworker immediately told us that this was a serious allegation. One prior report had also been for withholding medicine.

The AFST score was a Low risk protocol (i.e. a mandated screen out).¹¹ The caseworker thought the AFST score was too low. The caseworker said they would override the AFST Low risk protocol and screen the current referral in, because of the seriousness of the current allegation and the one prior referral with a similar allegation.

(2) Workers thought that the AFST over- and under-scored referrals specifically based on system involvement, i.e. welfare, public medical services, criminal history, or CYF history.

Interview findings: Six caseworkers and three supervisors said that the AFST over-scored families with more system involvement and under-scored families with less system involvement. For example, one caseworker said that families who do not use public welfare or medical services get scored lower than families who do: *“if you were poor and you’re on welfare, you’re gonna score higher than a comparable family who has private insurance. Because those people go to private therapists.”*¹² A caseworker said people who do not have a history of county involvement *“could totally be away from Big Brother forever.”* Workers’ beliefs align with prior work which suggest that the AFST is biased towards poor people and people with system involvement [119, 193]. We observed workers disregard the AFST score because they thought it was relying too heavily on system involvement and not taking relevant information into account.

Contextual Inquiry Observations: For example, we observed one caseworker review a report where they thought the AFST score was high because of system involvement, but they wanted to screen it out. A judge mandated that CYF look into the family after their child came to a juvenile probationary hearing. The child was currently incarcerated. The family had a number of prior referrals and used a lot of public behavioral health services (50+ times). Based on this history, this caseworker said *“I know [the AFST score] is gonna be high.”* When the caseworker ran it, the AFST score was a 20. However, the caseworker wanted to screen out the referral, because the child was already incarcerated, so it would be no use to investigate the family. The AFST over-scored this referral based on the family’s history and neglected highly-relevant context: that the child was in custody of the state and thus would not need to be investigated.

Overall: Prior work suggests that Black people have higher rates of poverty and system involvement, and that this accounts for other racial disproportionality in the Child Welfare system [108, 193]. Workers understood that the AFST was over- and under-scoring referrals based on system involvement and were correcting for that by disagreeing with the AFST’s recommendations for these referrals.¹³ Because system involvement is also correlated with race (Black families having more involvement than white families), this would account for the higher percentage of Black children than white children that the AFST-only decisions would have screened in but workers screened out (22.3% vs. 17.6%) and for the lower percentage of Black children than white children that the AFST-only decisions would have screened out but

¹¹This Low risk protocol was introduced in the AFST Version 2 [432]. It is analogous to the mandatory screen in label, but for screening out: so, the default is screening out, but it can be overridden by a supervisor.

¹²The official AFST FAQ documentation says *“receiving of public benefits”* did not necessarily increase a family’s AFST score: *“[F]or 45% of families, receiving of public benefits (e.g., SNAP, TANF)... was associated with lower scores than for similar families that did not receive those services”* [AFSTfaq]. As one supervisor pointed out, however, workers’ perceptions could still be right if there are *“other things that are affecting that score”* that are *“just more associated”* with public welfare records, e.g. public mental health records or criminal records.

¹³Prior work also suggests that CYF workers can reliably correct for limitations in the AFST [95].

workers screened in (15.9% vs. 29.6%) as seen in Figure 2.5. As stated in Section 2.12, this pattern of disagreement contributed to the 11% reduction in Black-white screen-in rate disparity from the AFST-only to worker-AFST decisions. In sum, because workers disregarded the AFST score more often when they perceived it to be over- or under-scoring based on system involvement (and not considering other relevant information), they were able to reduce racial disparities in call screening.

(3) Workers compensated for what they thought were the AFST's racial disparities caused by systemic racial biases in CYF reporting and county data collection.

Interview findings: Five caseworkers thought AFST-only decisions were racially disparate not only because the AFST over- and under-scored based on poverty and system involvement, but because of systemic racial biases in CYF reporting and in county information collected elsewhere—such as the medical system or the criminal system. These workers did not make screening decisions based on the race of the family, but they did consider the race of the family in order to account for systemic racial biases. These workers also thought that the AFST was biased because of over-reporting on Black families and systemic racism. One caseworker said that *“white people are not reported as much as Black kids”* and that they *“get a lot of reports on African-Americans and a lot of them are bogus.”* Another caseworker also agreed with this: *“I also think [the AFST is] very biased, but so is the world.”* This caseworker continued, *“the whole system is racially biased. ... It's the people entering the information [i.e. reporting families] that's affecting the [AFST] score.”*

Contextual Inquiry Observations: For example, we observed a caseworker review a report which included a fifteen-year-old boy who had not been to the dentist in five years and whose teeth were severely damaged: The caseworker said, *“he needs nine root canals, seven fillings”*. This was not reported by the dentist, however. The caseworker said that *“the dentist had all the information of the last five years of them trying to get [the boy] to come to the dentist... And [the dentist] didn't report him, because he's white.”*

Interview findings: Workers said they do not take the race of the family into account when making screening decisions. However, they do consider racial biases in CYF reporting and county information collection. For example, when asked if they take race into account when making screening decisions, one caseworker said *“we don't treat any of the cases differently.”* However, this caseworker later gave a hypothetical example of a referral where a Black family may get reported by *“people that don't deal on a daily basis with people of different cultures”* who *“automatically assume, like, ‘oh my God, holy shit, you can't swear at your kids like that.’”* However, this caseworker would understand that, because of the racial biases of the reporter and because the family is Black, the

However, this caseworker later gave a hypothetical example of a referral where a Black family might get reported by someone who doesn't *“deal on a daily basis with people of different cultures”* and who might *“automatically assume, like, ‘oh my God, holy shit, you can't swear at your kids like that.’”* But, this report would be unjustified: *“you and I may think [swearing at your kids is] tacky, but is that child hurt? Kid's not hurt.”* Another caseworker also said that they consider the race of the family in order to account for biases which could affect reporting:

“Colorblind assessment also feels like it's ignoring the point. So, I feel like [race is] definitely something that I take into consideration. ... Bias could be affecting the way that the information is being reported. ... I feel like I'm definitely more conscious of it now.”

Overall: Some workers understood that there were systemic racial biases that existed outside of their agency—in reporters and in professionals (e.g., in the medical and criminal systems) who create the data that they then make decisions based on. These workers expressed that they make screening decisions in order to compensate for these systemic racial biases. When it came to the race of the family, most of these workers did not consider race qua race, but rather race qua racial biases, which could color the information that they see in a report.

Collaborative decision-making

Screening decisions were not made by siloed, individual workers. Workers regularly made decisions collaboratively. They did so in two ways: First each report went through multiple (at least two) layers of workers in order to make the final screening decision. Second, workers often collaborated in ad hoc, impromptu ways, especially when workers were uncertain about a decision. We suspect that this collaboration contributed to mitigating workers' individual biases, and thus a reduction in the screen-in rate disparity in aggregate. However, this contribution is likely lessened, because final screening decisions are still made primarily by a few supervisors, and because the second kind of collaboration listed above is still informal and not applied consistently across all referrals.

Interview findings: First, as described in Section 2.10.1, the workflow on a standard referral necessitates that one caseworker and one supervisor review a referral. For most referrals, the caseworker receives a report, fleshes it out by searching the KIDS county database, makes a risk and safety assessment, makes a screening recommendation, and runs the AFST to generate a risk score. The report is then sent to a supervisor, who reads it in its entirety. One caseworker said the reports are not written by the caseworker alone—it is often a collaborative effort where the supervisor communicates with the caseworker about any inconsistencies, ambiguities, or mistakes in the report: *"it isn't like, 'I'm doing it. Clickety-click-click-click.'... [The supervisor] knows right away that I screwed something up."* If there are any holes in the report or any children in the report are younger than four, the supervisor may also ask for a *field screen* to send a field caseworker out to conduct a preliminary check on the family or gather more information about the report. Once the report is complete, the supervisor makes the final decision to screen the report in or out.

Supervisors primarily make the screening decisions. That being said, the process of both caseworker and supervisor reviewing a referral, communicating about the referral, making assessments of risk and safety, and making screening recommendations, is an important form of multi-layered, collaborative decision-making. Having the multi-layered process not only adds additional pairs of eyes to review each referral, it also ensures that the referrals are more likely to be reviewed by workers with different (demographic) backgrounds. One caseworker believed that the diversity among caseworkers was helping to reduce biases in the screening decisions: *"It's good to have different backgrounds with supervisors and others who make those decisions after it passes from our hands and it goes on to the next level of folks. And it's good to know there's diversity within those groups."*

Contextual Inquiry Observations: We also observed workers collaborating in ad hoc, impromptu ways that were not built into the standard decision-making process. Workers often discussed reports with other workers beyond the one caseworker and one supervisor assigned to the report. This occurred especially if they were uncertain about it. For example, while observing one caseworker (call them screener A), another caseworker (screener B) asked screener A to double-check that all the family

members on a referral were correct, since screener B had taken a call from this family previously. Supervisors often talked with each other (or sometimes caseworkers) if they felt uncertain about a screening decision. For example, if they felt uncertain about a report, one supervisor said, *“we’ll sit down and discuss it amongst the supervisors. ‘Well, what do you think we should do with this?’ It’s not like we all work in a little bubble.”* At one point, we observed a caseworker walk to a supervisor’s desk to discuss what to do about a referral that the caseworker had reviewed and the supervisor was making the final decision about.

This process was often informal and ad hoc: Workers worked in tightly-packed cubicles in a single room. So, they would talk to one another over the cubicles, often overhear one another talk about a report and add their two cents, or walk to another worker’s cubicle to ask them what they think about a certain referral. One supervisor described the call screening decision-making process as *“very collaborative”* and *“a group effort”*: *“Sometimes I’ll be reviewing and I’ll be like, ‘Hey, [supervisor], what do you think about this?’ You know, or we’ll just be talking about it in the room. If the reporting source calls back, someone else will hear it and be like, ‘Oh, that’s my reporting source’”*.

Overall: We suspect that this collaboration may have an effect of assuaging workers’ biases. Prior work suggests that workers’ biases play into the CYF screening process [193]. Some workers acknowledged that they have their own personal biases. For example, a caseworker said, *“I try to be conscious of my biases.”* With at least a caseworker and a supervisor reviewing each referral in detail, workers have more of a chance to correct for each others’ biases.¹⁴ For example, one caseworker said the CYF call screening process has *“a lot of checks and balances.”* Some workers themselves also expressed this view that multi-layered decision-making curbed biases. For example, when asked how workers make fair decisions, one caseworker said that *“it’s good to have... others who make those decisions after it passes from [caseworkers’] hands and it goes on to the next level of folks.”* By curbing workers’ individual biases, we hypothesize that such collaborative decision-making may have contributed to workers’ final screening decisions being less racially disparate in aggregate from 2016 to 2018. That being said, however, any bias-curbing effects of multi-layered, collaborative decision-making may be lessened due to caseworkers having less agency to influence the final screening decisions. For example, one caseworker said, *“I have very limited power... I don’t have the power of saying something’s not right.”* Furthermore, because the second form of collaborative decision-making described above was primarily informal, ad hoc, and not built into the standard decision-making for every referral, it is unclear how many referrals would have been affected by this form of collaboration.

2.13 How worker-AFST collaboration affected accuracy and accuracy disparities

2.13.1 AFST-only screening decisions were more accurate than workers when measured on outcomes that the AFST was trained to predict

So far, we have seen how workers made screening decisions which were less racially disparate than the AFST-only decisions. However, one of the primary arguments for the use of the AFST has been that it is more *accurate* than call screen workers [73,

¹⁴Here, we mean individual biases, such as implicit biases or overt prejudices. For systemic biases, such as those built into worker training or the bureaucratic process itself, having more eyes on a report is unlikely to curb them.

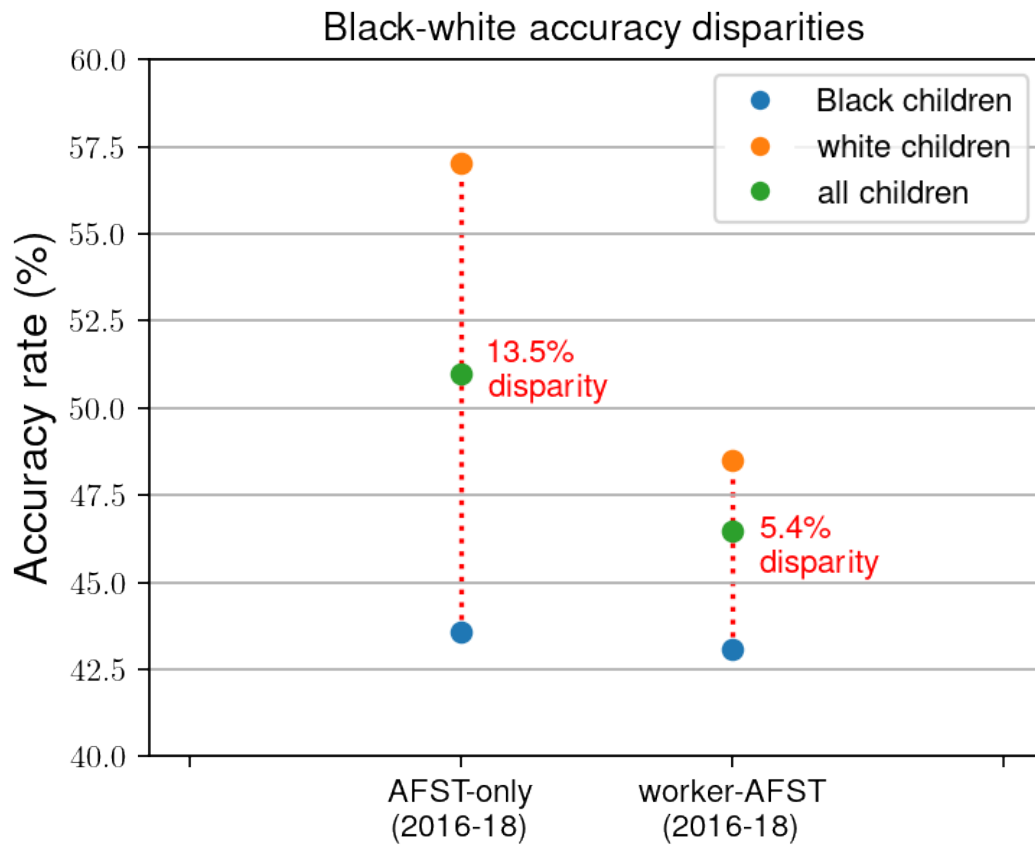


FIGURE 2.6: The differences in prediction accuracy between AFST-only and worker-AFST decisions.

88, 193, 393].¹⁵ For example, in the official Ethical Analysis of the AFST, Dare and Gambrill [88] claim that the AFST is “more accurate than any alternative” and argue that it “is hard to conceive of an ethical argument against use of the most accurate predictive instrument.” After all, an inaccurate screening decision could mean that a family is unjustly investigated—a possibly traumatic experience and a step too far towards child separation. Or it could mean that CYF does not intervene when a child will be harmed. In our setting, one concern is that a screening policy which makes less racially disparate decisions might make less accurate decisions [57, 287, 428]. For example, it may be the case that because workers make less racially disparate decisions than the AFST, they would also make less accurate decisions. In Figure 2.6, we see that this is ostensibly the case: the AFST-only screening policy would have been more accurate than the actual screening decisions made by workers from August 2016 to July 2018 (51.0% vs 46.5%).¹⁶ In Figure 2.6, we also see that the worker-AFST decisions were less racially disparate in terms of accuracy than the AFST-only decisions (5.4% vs 13.5%). In sum, although the worker-AFST decisions achieved lower prediction accuracy than the AFST-only decisions, worker-AFST decisions were also less racially disparate than the AFST-only decisions in terms of accuracy.

¹⁵On the other side, one of the primary arguments against its use has been that it is not accurate enough [70, 119].

¹⁶To reiterate, this AFST-only policy is defined as in Section 2.11 with a screening threshold of 15, i.e. screen in all High risk referrals and screen out everything else.

However, the definition of *prediction accuracy* we adopt is particularly important when interpreting these results. Accuracy is measured against outcomes that the AFST is trained to predict, i.e. re-referral and placement. Thus, a more precise interpretation of the results in Figure 2.6 is that the AFST-only was 4.5% more accurate than workers at predicting the outcomes that the AFST was trained to predict. However, prior work suggests that these outcomes are biased and contested [82, 95, 212]. Furthermore, as we will discuss in Section 2.13.2, our qualitative findings indicate that workers disagree with the use of these outcomes, and that they are predicting different outcomes than the AFST day-to-day. Therefore, although the AFST-only screening policy was better at predicting the outcomes it was trained to predict, it is important to keep in mind that the decisions made through worker-AFST collaboration may actually be better at predicting the outcomes that workers find useful or important for preventing child abuse or neglect.¹⁷

2.13.2 Qualitative: Workers and the AFST do not agree on prediction outcomes and accuracy measures

Workers believe re-referral and placement are bad proxies for risk of child abuse and neglect

In the past, CYF caseworkers have been asked about how confident they are in “*the AFST’s ability to accurately assess the risk of a future referral or out-of-home placement?*” [AFSTfaq]. Workers responded with lukewarm confidence because they thought the AFST was unable “*to take expected improvement or individual circumstances into account*” [AFSTfaq]. Yet, this may be the wrong question to ask entirely: workers may or may not be confident that the AFST can accurately predict the proxy outcomes that it was trained to predict. A more fundamental question is: are caseworkers confident that *any* assessments of risk of future referral or out-of-home placement will help them make better decisions to prevent child abuse and neglect?

Interview findings: Based on our observations and interviews with workers, we believe workers lean towards ‘No.’ Although the AFST was intentionally designed to complement workers judgment by nudging them to consider longer-term risk, in addition to short-term safety concerns [432], workers did not view the AFST’s target outcomes as relevant to their decision-making.

Re-referral. Five caseworkers and two supervisors thought that re-referral does not necessarily indicate risk of child abuse or neglect. One supervisor said that re-referral may not mean that the first decision to screen out was incorrect, since the second report could be for an entirely different reason than the first: “*it could have been referred because the mom was outside on the porch, smoking a cigarette with a newborn baby, and then it comes back mom beat the 10-year-old. I mean, it’s just a whole different reason.*” Three caseworkers and one supervisor said that some reporters misuse the system and report families for unnecessary reasons. For example, one caseworker gave “*retaliation reports*” as an example: “*I call on you. You call on me. I call on you.*” In our contextual inquiry, we observed a caseworker review a referral where a divorced couple called on each other three different times and each report was unsubstantiated: the caseworker said these were likely false retaliation reports. The designers of the AFST themselves noted problems with using re-referral as a target outcome and they removed it from the AFST model in 2018 [432].

¹⁷Here, we hearken back to a question brought up in the *Ethical Analysis* of the AFST: “*The question is, how can we make the fewest errors in our efforts to protect children and families?*” [88]

Placement. Since August 2018, the AFST (Version 2) only predicts placement. However, four caseworkers and one supervisor also found issue with using placement in foster care as a proxy for child abuse or neglect. A caseworker said, “*knowing the risk of removal within two years is not feeling like it’s super relevant to the decision that is needed. And it is very little to do with immediate safety or anything like that.*” Two caseworkers said that children were often placed in foster care without any concerns of child abuse or neglect: there are often other reasons for placing a child. One caseworker said:

“The majority of these cases [when children are placed in foster care] are not child abuse in nature, it’s parent-child conflict, the kid doesn’t want to live with mom or dad or the grandma, the child is saying, ‘I don’t feel safe, I don’t want to go home.’ And if they go to the police and then saying that information, some police will take custody of that child and the court has to place that kid. So it’s just a lot of other variables going on that decide whether or not this child is going to be placed.”

On the other side, one caseworker and one supervisor said that placement is not the right option for many families who do have concerns of child abuse or neglect. One caseworker said families are often reported for safety concerns, without any possibility of placing a child in foster care: “*just because the report is being made doesn’t mean that a kid is going to go into placement.*” In prior work, De-Arteaga, Dubrawski, and Chouldechova [94] similarly note that “[n]ot all cases where there is a risk to the child result in out-of-home placement.” Lastly, another caseworker said that placement was not a good outcome to measure because it was a kind of self-fulfilling prophesy: “*risk of removal in two years is inherently going to be increased by our [CYF] involvement, because we’re the only ones that can remove the children.*” De-Arteaga, Dubrawski, and Chouldechova [94] and Coston et al. [82] also suggest this point.

Workers did not make decisions based on risk of re-referral or placement

Not only did workers say that re-referral and placement were not helpful outcomes to predict, we also observed workers making screening decisions based on assessments of entirely different outcomes.

Contextual Inquiry Observations: We observed that call screen workers did not try to predict whether or not children in a referral would be re-referred or placed in foster care within two years. Workers’ screening decisions were based on different, shorter-term outcomes related to child safety.

For example, while reviewing referrals, one caseworker said that caseworkers made decisions by looking for “*safety concerns.*” This caseworker looked for the following, for example: signs that the child felt unsafe, would be hurt, was left home alone a lot, was in contact with a child molester, or that the caregiver was under the influence of drugs, etc. The caseworker also said that caseworkers asked reporting sources about the following, for example: *Is there food in the fridge? Do the kids have sheets on their beds? Is there furniture in the home? Do the parents have drug or alcohol problems? Mental health concerns? Domestic violence concerns?*

Clearly, re-referral or placement were rarely among the factors that caseworkers assessed for. Furthermore, these factors were not indicative of a risk of child abuse or neglect over the span of two years: it will not take a child two years to starve if there is currently no food in the fridge. The factors that workers looked for when making their screening decisions were either shorter-term safety concerns, or specific details

about the referral that could be longer-term sources of risk, regardless of whether they would lead to removal from home.

2.14 Discussion

2.14.1 Interpretation of the findings

Automated AFST screening decisions would have yielded larger racial disparities

In this paper, we compared two policies for using the AFST in CYF call screening from August 2016 to May 2018: first, the *hypothetical* policy where the AFST entirely automates screening decisions; and second, the *actual* policy where workers use the AFST to inform their decisions. We evaluated these policies in terms of racial disparities in both screening rates and predictive accuracy. Our results in Section 2.12 suggest that the automated AFST-only policy would have resulted in larger disparities in call screening: the AFST-only Black-white screen-in rate disparity would have been 20.1%. This is larger than both the 11.3% pre-AFST disparity from January 2015 through July 2016 and the 9% disparity for AFST-assisted worker decisions from August 2016 through May 2018. Had the AFST automated screening decisions at CYF, our results suggest that this would have increased the disparity in the rates at which Black versus white children were screened in for investigation.

Instead, when workers used the AFST, this disparity of 20.1% decreased to 9%. In order to understand why this occurred, we first looked at how workers disagreed with the AFST. Rather than disagreeing with the AFST in ways that directly reduced disparities, we found that workers disagreed with the AFST across the board. Some of workers' disagreements contributed to greater disparity and some served to reduce disparity. Yet, overall, worker-AFST disagreement led to a reduction in screen-in rate disparity. We then conducted a contextual inquiry and interviews with call screening workers at Allegheny County CYF to further understand this pattern. As shown in Section 2.12.2, we found that workers were not surprised that the AFST gave racially disparate screening recommendations. Workers made holistic screening decisions based on their knowledge of relevant context beyond just the AFST score, and they believed that this ultimately led them to make less racially disparate screening decisions than the AFST alone would have. Workers made decisions on a case-by-case basis, focused primarily on the risk and safety of the children involved in each individual report. Within these bounds, some workers made a conscious effort to reduce unwarranted racial disparities. Others believed that any reduction in disparities that occurred due to their disagreements with the AFST was incidental and unintentional, viewing this as a side effect of making decisions holistically based on various sources of information available to them. However, the effect of reducing disparity in aggregate was the same.

Our results run counter to much prior work on racial disparities in human-AI systems. As Green [144] writes, the "vast majority of research suggests that people are unable to provide reliable oversight of algorithms" and that human discretion worsens racial disparities. For example, judges have disregarded criminal risk assessment recommendations in ways that worsen racial disparities [7, 85, 196, 399, 400, 401, 447]. In Mechanical Turk experiments, people have been shown to use their discretion to make racially biased decisions [146, 148]. We present a case study in the opposite direction: worker discretion served to *decrease* racial disparities in the AFST from August 2016 to May 2018, echoing De-Arteaga, Fogliato, and Choudhova [95], who suggest that CYF workers used their discretion to counteract errors

in the AFST. As discussed below, further research is needed to understand the impacts of worker-AFST decision-making on predictive accuracy, and to untangle the extent to which worker discretion served to decrease unwarranted versus warranted screening disparities.

In the context of this prior literature, our findings suggest that caution is needed when drawing broad generalizations about the impacts and dynamics of human-AI decision-making (cf. [95, 170, 417]). In practice, the interplay of human and AI judgment may yield very different results depending on the specific domain under study, the abilities and expertise of the humans involved, and the organizational contexts in which humans and AI are making predictions and decisions.

Was the AFST responsible for the reduction in disparity from pre- to post-AFST?

A key concern for any policy change in child welfare is whether it will reduce or amplify existing racial disparities. In the description of the official AFST Impact Evaluation [140], Allegheny County Department of Human Services suggests that the “AFST led to reductions in disparities of case opening rates between black and white children” [190]. We also observed a reduction in screen-in rate disparity from before the introduction of the AFST to after (11.3% disparity from January 2015 through July 2016 to 9% disparity from August 2016 through May 2018). However, it is not clear that this reduction in disparity was caused by the introduction of AFST in August 2016. It is difficult to determine whether and to what extent changes in disparities can be attributed to the AFST, since the AFST was implemented across all Allegheny County CYF screening decisions (without A/B testing or randomization) and there were a number of external factors which influenced call screening—such as state-wide amendments to mandated reporting laws at the end of 2014 or any screening policy changes internal to CYF.¹⁸ It is also possible that the kinds of referrals being referred from August 2016 through May 2018 were different from those in January 2015 through July 2016, regardless of the AFST’s recommendations.

Our results show that the AFST gave recommendations which were more racially disparate than workers pre-AFST. So, several interpretations are possible. For instance, it may be that the workers reduced the overall disparity on their own, by making decisions largely as they had prior to the introduction of the AFST. On the other hand, if the introduction of the AFST did contribute to a reduction in screen-in rate disparities from pre- to post-deployment, our results suggest that workers’ discretion was integral to this reduction. One possible explanation is that the introduction of the AFST may have led workers to reflect on their own biases, resulting in decisions that were less racially disparate than their decisions in the year and a half prior to the introduction of the AFST. A second possibility is that workers and the AFST have complementary strengths and biases, and that the interplay of these led to less biased screening decisions overall. These possibilities are neither mutually exclusive nor exhaustive: some combination of these mechanisms may be at play, and additional mechanisms may be possible beyond those mentioned here. We leave it to future work to further explore whether and how the introduction of the AFST may have impacted disparities from pre- to post-deployment. In any case, our results suggest that the observed reduction in screen-in rate disparity can be attributed to human workers, whether or not they used information from the AFST to do so.

¹⁸Here, we refer to amendments to the Pennsylvania Child Protective Services Law (CPSL) that went into effect on December 31, 2014.

Was the AFST more accurate than workers?

Prior work suggests that the AFST makes more accurate predictions than workers [73, 88, 193, 393]. While our results in Section 2.13.1 show that the accuracy disparity for AFST-only decisions would have been higher than that of worker-AFST decisions, they also show that AFST-only decisions would have been more accurate than worker-only decisions from January 2015 to July 2016 (55.2% vs 49.7%) and worker-AFST decisions from August 2016 to May 2018 (51.0% vs 46.5%).¹⁹ This increased accuracy is one of the primary reasons for the use of the AFST [73, 88, 193, 393]. An official Human Services [192] statement echoed that “*not using [the AFST] might be unethical because of its accuracy.*” Yet, the AFST V1’s accuracy is measured in terms of specific proxies for abuse or neglect (re-referral and placement) over a long time range (two years). In Section 2.13.2, we found that workers did not agree with these choices of proxies, and that they viewed themselves as predicting immediate safety risks—in other words, a *different outcome* than the AFST, over a *shorter time span*. This is consistent with findings from prior work [91, 94]. Thus, while our results suggest that the AFST is better at predicting the proxy variables it was trained to predict, it remains an open question how the AFST compares at the prediction and decision tasks that workers are actually trying to perform. Vaithianathan et al. [435] suggest that higher AFST V2 risk scores identify children with higher rates of injury-related hospitalizations, which may be closer to the kinds of target outcomes that workers actually consider when making decisions.²⁰ However, our results suggest that workers predict different outcomes depending on the specific referral, many of which may not involve child hospitalization. Overall, our results in Section 2.13.2 complicate the argument that the AFST is more accurate than call screen workers, and point to critical directions for future research. Without first understanding workers’ objectives when making predictions and decisions, such accuracy comparisons may essentially be evaluating workers’ performance on a game that they are not playing.

2.14.2 Design Implications

Our results suggest that worker-AFST collaboration yielded lower racial disparity in screen-in rate than AFST-only decisions, and lend insight into how this reduction occurred. Here, we provide design implications for potential ways to improve human-AI collaborative decision-making in child welfare and related contexts.

- **Explainable AI and interfaces to empower workers in mitigating algorithmic limitations.** Our findings suggest that workers’ ability to identify instances where the AFST was over- or under-scoring a referral may have contributed to a reduction in racial disparity (2.12.2). However, it remains unclear how accurately and reliably workers are able to do so. Thus, one possible design implication is to develop interfaces that assist workers in identifying specific instances where an algorithm may be more or less reliable. For example, in the AFST context, providing local explanations for AFST recommendations may help workers in calibrating their reliance on the tool, case-by-case. However, given recent empirical findings demonstrating ways such local explanations can backfire in ways that harm human-AI decision-making (e.g., [20, 329]), further research is needed to understand how these can be designed and presented effectively.

¹⁹We measure the AFST’s accuracy differently than most prior work, which uses AUC (area under the receiver operating characteristic curve).

²⁰Note that AFST V2 also predicts foster care placement within 2 years, but not re-referral.

- **Value sensitive AFST models.** Our results suggest that workers often adjust for what they perceive as limitations of the algorithm. For example, workers believed that the AFST was unable to sufficiently account for certain features (e.g., details of the actual allegation), and the outcomes predicted by the AFST did not align with the outcomes that workers were predicting. Zhu et al. [474] suggest involving stakeholders into the design process to make sure their values are incorporated into the design of an algorithmic system from the beginning. One design implication is for AFST designers to engage with child welfare workers, among other key stakeholders, and to incorporate their insights and feedback in future iterations of AFST. Some value misalignments between workers and the AFST were intentionally designed into the system, with the goal of nudging workers towards different practices. However, in practice we observed that these misalignments meant that workers often needed to work around the algorithm instead of working with it.
- **Promoting more collaborative decision-making.** Child welfare workers often made decisions collaboratively, both formally and informally. We suspect that this may have had the effect of curbing workers' individual biases. However, in formal collaboration, caseworkers felt they had little agency—even though they made screening recommendations. The first design implication is to encourage more regular conversations between caseworkers and supervisors about caseworkers' recommendations, so they do not get overlooked. There may also be opportunities for the AFST interface to promote more informal collaboration. For example, future versions of the AFST or similar tools could include a feature to suggest caseworkers in the office that have dealt with similar referrals in the past. This could enable workers to collaborate with the right person on each referral, which may be particularly helpful when dealing with highly uncertain referrals.
- **Diversity among child welfare workers.** Child welfare workers were generally aware of their own individual biases. Another way to curb these biases may be to increase diversity among the child welfare workers, especially supervisors (who make the final decisions). As one caseworker stated, it is important to ensure that there are screeners and supervisors who have "*lived the experience*" of the families in a report.
- **Be cautious about automating decisions in contexts with existing racial disparities and biases.** Our results suggest that workers reduced racial disparities in the AFST through their patterns of disagreement with the AFST score. Based on our findings, we strongly caution against fully automating decisions in high-stakes, real-world social contexts. This recommendation aligns with prior work in the child welfare context [95]. For the same reasons, even partial or 'soft' automation (like the mandatory screen-in policy in the AFST context) should be approached cautiously to ensure that they do not incentivize against potentially *productive* forms of human disagreement with algorithmic recommendations.

2.14.3 Limitations & Future Work

One limitation of our methods is that the data used in our quantitative analysis is from August 2016 through May 2018, whereas the contextual inquiry and interviews were conducted in July 2021. Thus, if workers changed how they used the AFST

significantly or they are unable to remember how they used it previously, then our results from the contextual inquiry and interviews may not reflect the reasons why workers reduced disparities from 2016 to 2018. This is likely not entirely the case, since caseworkers' workflow has remained largely consistent over this time frame (even though there have been some changes in CYF policy around the use of the AFST, including the roll-out of a new V2 model).

Another limitation of the current work is that we focus on the impacts of worker-AFST decision-making on overall disparities. We do not investigate the extent to which worker discretion served to decrease *unwarranted* versus *warranted* disparities in the AFST's screening recommendations. Our results demonstrate that worker-AFST decision-making served to reduce the disparity in child maltreatment screening rate between Black and white children compared to algorithm-only decisions. As discussed, this is an interesting finding in itself, in light of prior empirical research suggesting that human discretion often increases rather than decreases disparities. This finding also has practical implications, as racial disparities in child maltreatment screening have real consequences for children, families, and communities. For example, higher screen-in rates indicate higher levels of state intervention into families' lives, starting with an investigation. Such disparities may indicate uneven application of interventions or investigations. Importantly however, higher disparities in screening, on their own, do not necessarily imply unfairness. Disparities may be warranted if they reflect genuine differences in underlying distributions: a higher screen-in rate for one group may be justified if there is higher need or higher risk of maltreatment among children in that group [100, 109, 122, 230]. For this reason, naively attempting to equalize screening rates across demographic groups without regard for actual children's needs and safety could be harmful and counterproductive. However, it is clear from our interviews and contextual inquiries that this is *not* how workers reduced disparities in algorithmic decisions. Rather, workers appeared to reduce disparities by making holistic assessments of child risk and safety based on all of the information available to them, and by working to mitigate limitations that they perceive in the algorithm. An important direction for future research is to untangle the extent to which worker discretion serves to decrease unwarranted versus warranted screening disparities. To support such investigations, it will be critical to overcome limitations of current approaches for measuring accuracy, discussed in Sections 2.11.3 and 2.13.1.

We present two additional avenues for future work, informed by limitations of our current work. First, we leave it to future work to conduct a more comprehensive evaluation of racial disparities between worker-only and worker-AFST screening decisions. We did not focus on these results in the current paper, due to several confounding factors that complicate the analysis. These included changes in external factors—such as a 2015 state-wide policy change in mandated reporting laws and the COVID-19 pandemic since at least March 2020—for affecting the County reporting rates and the CYF screening process.²¹ Second, as discussed, measuring prediction accuracy is limited by the use of proxy outcomes (re-referral or placement) that do not align with the outcomes that humans are predicting [45, 95]. It is possible that people are better than algorithms at the prediction task they are actually performing. While beyond the scope of the current paper, this remains a critical direction for future work. For example, future research could involve better understanding what constructs human workers are predicting, operationalizing these as target measures,

²¹We visited Allegheny County CYF's Intake Department in July 2021 when many call screen workers were working from home, due to the COVID-19 pandemic. This has had an impact on the CYF call screening workflow since March 2020.

and then re-running some of the accuracy comparisons presented in this paper using those measures.

Worker-AI Interaction, Part 3

Exploring Worker-AI Interface Design Concepts

This chapter is taken from the following paper:

Anna Kawakami*, Venkatesh Sivaraman*, Logan Stapleton, Hao-Fei Cheng, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. "Why Do I Care What's Similar?" Probing Challenges in AI-Assisted Child Welfare Decision-making Through Worker-AI Interface Design Concepts." In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*.

Abstract. Data-driven AI systems are increasingly used to augment human decision-making in complex, social contexts, such as social work or legal practice. Yet, most existing design knowledge regarding how to best support AI-augmented decision-making comes from studies in comparatively well-defined settings. In this paper, we present findings from design interviews with 12 social workers who use an algorithmic decision support tool (ADS) to assist their day-to-day child maltreatment screening decisions. We generated a range of design concepts, each envisioning different ways of redesigning or augmenting the ADS interface. Overall, workers desired ways to understand the risk score and incorporate contextual knowledge, which move beyond existing notions of AI interpretability. Conversations around our design concepts also surfaced more fundamental concerns around the assumptions underlying statistical prediction, such as inference based on similar historical cases and statistical notions of uncertainty. Based on our findings, we discuss how ADS may be better designed to support the roles of human decision-makers in social decision-making contexts.

2.15 Introduction

From social work to content moderation to education and legal practice, data-driven AI systems are increasingly used to augment human *social decision-making*: decisions that rely upon inferences and predictions about the intentions, behaviors, and beliefs of other people [23, 260]. AI-based tools are commonly introduced into social decision-making settings with the promise of overcoming human limitations and biases [235, 439]. However, in such settings, AI judgments are themselves likely to be imperfect and biased, even if in different ways than humans [11, 119, 177]. Prior work across a range of social decision-making contexts indicates potential for front-line decision-makers and ADS tools to complement each other's capabilities and help to overcome each other's limitations (e.g., [61, 96, 169, 213]). Yet, achieving such synergy in AI-augmented decision-making is still far from guaranteed. Recent research demonstrates ways that AI augmentation can backfire in practice, harming decision quality compared with either human or AI-based decisions alone (e.g., [147, 329, 401, 416]).

To date, little is known about how to design ADS interfaces to support effective AI-augmented social decision-making in real-world contexts. Such domains tend to invalidate assumptions often made in less socially-fraught decision settings: that the available data features are sufficient for optimal decision-making, that the desirability of various decision outcomes is well-established and agreed-upon, and that the outcome of each individual data point is inherently predictable. By contrast, in social decision-making contexts, human decision-makers must regularly operate on nuanced and unstructured knowledge, incorporate subjective and potentially contested values into decision-making, and make judgments despite high levels of irreducible uncertainty about people's future behavior or life circumstances [96, 221, 303]. Standard approaches to implementing ADS are often at odds with these complexities. For example, the quantitative accuracy of an algorithm may not capture that so-called "ground-truth" labels are socially contested, as has been observed in toxicity detection [141] and child maltreatment screening [96]. As a result, the growing body of AI design knowledge obtained from less contentious and value-laden decision settings [250] provides only a partial understanding of human-AI complementarity in social decisions.

In this paper, we begin to explore how these complexities manifest through the perspectives of frontline decision-makers themselves. To probe the challenges frontline workers face in their day-to-day interactions with an ADS, we consider potential re-designs or augmentations to the worker-AI interface for the Allegheny Family Screening Tool (AFST), an ADS which has been in active use for nearly half a decade. The AFST is intended to help child maltreatment hotline workers manage high volumes of referrals, assess their risk, and prioritize cases for further investigation [74, 436]. As we discuss in [221], however, workers face a range of challenges in using the ADS, including limited insight into the algorithm, value misalignments between the algorithms' objective function and their own, and organizational pressures around reliance on the tool. Overall, although the ADS has been in use for several years, many workers continue to perceive its current design as a missed opportunity to effectively complement their own abilities. In parallel with the inquiries presented in [221], therefore, we sought to elicit workers' perspectives on how specific re-designs or augmentations to the ADS interface could better support their daily work.

Drawing from prior literature as well as workers' ideas and feedback, we created a range of design concepts to probe workers' challenges and concerns around

the AFST, and their desires for new kinds of worker-AI interactions or more fundamental changes to the underlying ADS and its role in decision making. Through two rounds of design interviews with a total of twelve call screeners and supervisors at Allegheny County, we observed strongly favorable reactions to concepts that promoted worker understanding, control, and feedback—though workers’ notions were not always aligned with those discussed in prior human-AI interaction and interpretability literature. At the same time, our design concepts served to open conversations that revealed discomfort around underlying assumptions of statistical prediction, such as case comparison and statistical notions of uncertainty. Based on our findings, we discuss how ADS may be better designed to support the roles of human decision-makers in social decision-making contexts.

2.16 Background and Related Work

2.16.1 Designing interfaces to support AI-assisted decision-making

ADS are increasingly used to support human work in deeply social contexts, such as social work, education, healthcare, and criminal justice (e.g., [74, 175, 321, 416]). To ensure that these systems do more good than harm, it is critical that they are designed to bring out the best of human ability while also helping to overcome human limitations. To date, scientific and design knowledge remains scarce regarding how the strengths of human and AI judgment can be effectively combined in practice, while mitigating the drawbacks of each [20, 96, 169]. Empirical results from research studying AI-augmented decision-making have been mixed. A long line of literature in human-computer interaction (HCI), human factors, psychology, and management science demonstrates that humans are often either too skeptical of useful AI predictions or too dependent upon erroneous predictions and recommendations (e.g., [45, 102, 147, 258]).

Recent research provides early evidence that empowering human workers to evaluate and (as appropriate) second-guess ADS predictions may support more effective and equitable decision-making (e.g., [61, 96, 169]). Recently, researchers have proposed new support tools and interfaces to improve AI-assisted decision making, often by providing explanations of AI behaviors. However, several recent experimental studies have shown that, contrary to researchers’ intuitions, presenting in-the-moment explanations for an algorithm’s prediction can backfire, encouraging humans to over-rely on algorithmic outputs even in the presence of large errors that they may otherwise have noticed (e.g., [bandsal2021does, 329]). By contrast, some recent work signals potential for relatively simple cognitive cues or reflection prompts to help foster more effective use of ADS [46, 175, 329]. It remains an open research question how best to foster human–algorithm synergy, particularly in real-world AI-assisted decision-making contexts.

2.16.2 Challenges in ADS for social decision-making

Recent research has begun to explore the design of ADS to assist human decision-makers in complex social decision-making contexts. Following [260, 303], we use the term “social decision-making” to refer to tasks in which a decision-maker must reason or make predictions about the intentions, beliefs, and behaviors of other people—for example, choosing candidates for a job or a loan, or assessing the risk of future crime or child maltreatment. Although these decisions are often structured and systematized by the public institutions that enforce them (including through the use of

high- or low-tech algorithms [360]), prediction about the future behavior of individuals often resists clear-cut decision rules.

The standard approach to developing ADS often assumes the existence of a “ground truth” by which an optimal prediction function can be calculated. For example, a physician’s accuracy in detecting abnormalities in medical images can be measured by comparing their decisions to an expert consensus [322, 424]. By contrast, the ground truth labels in many social decision-making tasks—for example, whether an observed behavior is considered socially “harmful”—may be socially contested [141, 235] or represent an imperfect proxy for what decision-makers actually predict [126, 202]. Moreover, risk predictions are often highly uncertain due to individual variability, an abundance of unmeasured causal factors, and long time-frames over which outcomes are measured [272, 303, 353]. These features motivate the consideration of AI-assisted social decision-making as a subcategory of AI-augmented decision-making, requiring greater attention to the sociopolitical and cognitive contexts within which these algorithms function [221, 360].

In the current study, we focus on the context of AI-augmented social work, a high-stakes social decision-making setting where the use of ADS tools is rapidly spreading [74, 354, 361, 475]. Recent work has begun to investigate how to design interfaces to support more effective AI-augmented decision-making in social work contexts. For example, to help workers assess risk of child maltreatment, Zytek et al. [475] developed and evaluated an AI explanation dashboard to help social workers understand how an ADS arrives at its predictions. Our study complements and extends this prior work by broadening the scope of possible interface designs to capture other information the ADS may be able to provide, as well as other possible roles the ADS could play in social workers’ day-to-day work. In the following section, we provide a brief overview of child maltreatment screening, the specific decision task that forms the basis of our study.

2.17 Study Context

2.17.1 The Allegheny Family Screening Tool (AFST) and Workplace Context

In child welfare agencies across the United States, call screeners and their supervisors make screening decisions on a large volume of child maltreatment referrals every day, reviewing an extensive amount of administrative history for each case. Making systematic decisions in such complex, high-stakes settings is extremely challenging, requiring a balancing act between “erring on the side of child safety” and “erring on the side of family preservation” [74, 364]. To aid workers in making more systematic and efficient screening decisions, child welfare agencies have begun to introduce new ADS tools to support call screeners’ and supervisors’ daily decision-making.

We focus on the **Allegheny Family Screening Tool (AFST)**, an ADS introduced to the Allegheny County, Pennsylvania’s **Office of Children, Youth and Families (CYF)** Intake/Call Screening Department in 2016. The AFST assists child protection hotline call screeners in prioritizing and assessing risk in child maltreatment cases, by providing a risk score indicating the likelihood that a child may be placed out of home in two years. In Figure 2.7, we show a screenshot of the current interface showing an AFST score: For every case, it outputs a single risk score between 1 (low risk of future placement, left-most side of the bar) and 20 (high risk of future placement, right-most side of the bar) [436]. In Figure 2.8, we overview how the AFST is



FIGURE 2.7: A screenshot showing the AFST interface that the call screeners and supervisors see when using the tool [433]. The AFST risk score outputs a single number between 1 (low risk of future placement, left-most side of bar) and 20 (high risk of future placement, right-most side of bar) for every case. In the image, the tool gave a risk score of 5.

integrated into CYF’s screening and investigation process: First, a caller (such as a teacher) calls the CYF hotline center to make a potential child maltreatment report, i.e., a **referral**. A call screener must then make a recommendation about whether to screen in or screen out the report for investigation. To make this recommendation, the call screener considers the following information: Current allegation details based on the caller’s claims, other information sources from public records, and the AFST score (which is generated after the call screener gathers relevant administrative and history information). In their recommendation, the call screener may agree or disagree with a low or high AFST score. After the call screener gives a recommendation, the case report, the AFST score, and the screener’s recommendation are passed to the call screener’s supervisor, who makes a final screening decision. The supervisor may decide to agree or disagree with the call screener’s recommendation and/or the AFST score. However, if the AFST score is 18 or higher (a **mandatory screen-in score**) and the supervisor decides to screen out the case, they must go through an **override process**. Screened-out cases are not escalated into the child welfare system but are recorded for future consideration. Screened-in cases are assigned to caseworkers, who may further observe, investigate, or intervene with the family. For cases in which the supervisor believes there is insufficient evidence towards both decisions, the supervisor may ask the call screener to gather additional information (e.g., by calling the reporting source) before making their final decision.

2.17.2 Prior Studies on the AFST

The AFST is one of the most well-known public sector ADS, examined extensively by researchers in the HCI and ML communities [64, 74, 96, 429]. Other public sector agencies are beginning to view the AFST as an example of what ADS in social work could or should look like [354]. However, most existing studies on the impact of the AFST on workers’ screening decisions are based on analyses of historical, retrospective administrative data [61, 96]. To date, no prior studies have explored how the AFST might better support the call screeners and supervisors who interact with the tool every day. In fact, until recently, there was minimal insight into how call screeners and supervisors use the AFST even in its current form.

In [221], we conducted field studies at Allegheny County’s CYF to understand how call screeners and supervisors calibrate their reliance on the AFST. Overall, we found that workers faced several challenges with understanding and integrating

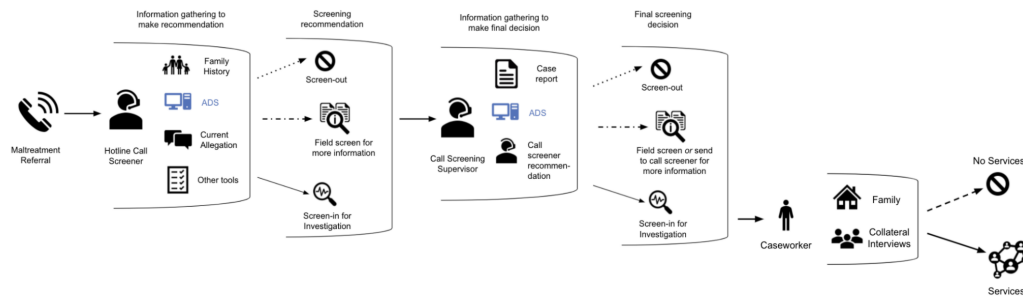


FIGURE 2.8: Diagram showing a high-level overview of Allegheny County's child maltreatment screening process, illustrating when an ADS assists call screeners' and supervisors' screening decisions. Taken from [221].

the AFST risk score into their decision processes. Workers felt they were provided little training and information on how the model works, and they grappled with value misalignments and organizational pressures that impacted their reliance on the score. These findings motivate the present design research study, in which we further examine how workers could be better supported in using an ADS, including (but not limited to) ways to understand and calibrate reliance on the AFST. We further detail how the findings from [221] inspired specific design concepts in Section 2.19.

2.18 Methods

To understand call screeners' and supervisors' current challenges, desires, and underlying needs for additional support in using the AFST, we visited Allegheny County's Office of Children, Youth and Families (CYF) and conducted a series of field observations, semi-structured interviews, and design research activities [221]. This paper presents findings from the latter design activities, which were conducted over two site visits across a period of two weeks to allow time for iteration between visits. We generated ten design concepts, each envisioning different ways of augmenting or redesigning the AFST (cf. [90, 177]). Before the first visit we prepared an initial set of three design concepts, largely focused on notions of uncertainty and comparison with historical cases. Building on our observations and design concept-related discussions from the first visit, we then expanded the set of design concepts to include a broader scope of possible redesigns or interface augmentations. The ten resulting design concepts can be organized into four broad categories, as described in Section 2.19 and visually summarized in Figure 2.16.

In each visit, we provided participants with an overview of the study purpose and methods, and obtained their consent for participation and audio recording or note-taking. Twelve participants consented to audio-recording and one consented to having notes taken. Given that we conducted these study sessions during participants' working hours, some participants were unable to complete all study activities. In total, we interviewed 12 call screeners and supervisors. The first visit included five call screeners and two supervisors, and the second visit included six call screeners and four supervisors. Two of the call screeners and two of the supervisors were present during both the first and second visits. During the second

visit, these four participants were only presented with newly-generated design concepts, which they had not already seen during the previous visit. Table D.1 shows participant demographics, presented in aggregate form to avoid making individual workers identifiable within their workplace.

2.18.1 Design concept generation

During the first visit, we interviewed workers and observed them as they made AI-assisted decisions to identify their challenges and needs for additional support [221]. Following the first visit, our research team conducted Interpretation Sessions [184] to identify design opportunities to support more effective and responsible AI-assisted decision-making with the AFST. The study protocol used in our second visit included prompts around a total of ten design concepts, including concept sketches to illustrate seven of these concepts visually.

During the second visit, we explored each design concept with participants via multiple prompts, following prior work exploring the design of ADS tools [103, 175]. For example, we asked workers to envision how they might use a given design concept, whether and how they thought it might influence their decision making, how useful or promising they find the design concept overall, and what alterations they might wish to make to the presented concept. With three of these concepts (**Q&A with the AI**, **Type of “risk” predicted by worker vs AI**, and **Measuring quality of worker-AI decisions**), our primary goal was to support design ideation rather than evaluation. Therefore, for these concepts, we did not ask questions about usefulness. Beyond the pre-determined prompts for each concept, we flexibly probed on other interesting topics that emerged in participants’ responses.

Many of the design concepts in our study were made more concrete by showing example case data, intended to help participants envision themselves using each interface. For the concepts presented in the first session, we adapted examples from real de-identified historical cases, and applied appropriate algorithmic techniques to generate uncertainty information and nearest neighbors. For example, the **Similar past case outcomes** concept was produced by computing nearest neighbors of each case using a human-interpretable re-encoding of the AFST’s input features. The **Uncertainty due to unusualness** visualization was created by comparing the number of these nearest neighbors that fell within a particular distance threshold. For the design concepts generated between the first and second visits, we wrote hypothetical vignettes and data features guided by prior analyses of the historical data. While these details were not the focus of the study, we took note when participants found these examples unrealistic as opportunities to improve future designs.

2.18.2 Analysis

Following [103, 177], we analyzed participants’ feedback on each design concept separately, while also examining trends in their overall feedback across concepts and iterations (see Figure 2.16). We collaboratively worked through 9.5 hours of transcribed audio recordings and six pages of additional notes. For each design concept, we focused our analysis on (1) how useful participants expected the concept would be, (2) underlying reasons why participants expected that a given concept would be more or less useful, (3) specific ways participants imagined they could make use of the presented concept in their day-to-day work, and (4) any suggestions for design modifications. To help in better understanding participants’ feedback, we

triangulated their responses with our observations of how they currently make AI-assisted decisions with the AFST.

2.18.3 Ethics and Participation

Throughout the study, we reminded workers that their participation was completely voluntary and that their responses would be kept anonymous. To prevent workers who participated in the study from being identified within their workplace, we only report participant demographics at an aggregate level.

2.18.4 Positionality

We acknowledge that our experiences and positionality shape our perspectives, which guide our research. We are all researchers working in the United States. Our academic backgrounds range across interdisciplinary fields within Computer Science, including HCI and AI. Some of us have prior experiences studying social work contexts or other public-sector decision-making contexts in the United States but not elsewhere in the world. None of us have been investigated by a child welfare agency nor adopted or involved in the foster care system. In addition, none of us have professional experience in child welfare. All authors except two live in Allegheny County; the other two live in Minnesota and California. To conduct this research, we collaborated with Allegheny County’s Child, Youth, and Families Department as external researchers. The analysis and writing were conducted independently from the department.

2.19 Design Concepts

In this section, we briefly introduce our ten design concepts, seven of which had visual representations (see Figures 2.9-2.15). For clarity of presentation, we group the concepts across four broad design topics: **Discrepancies between human and AI decision-making**, **Enabling worker-AI dialogue**, **Communicating uncertainty**, and **Measurement and feedback on historical decisions**.

2.19.1 Discrepancies between human and AI decision-making

In the absence of formal insight into how the AFST works, workers have improvised strategies to learn about the AFST’s capabilities and limitations on their own (e.g., by systematically testing out different outputs to the model and examining the outputs). They desired opportunities to gain more direct information about the model, and particularly to better understand how and why the AFST’s judgments differed from their own [221]. Therefore, we explored three design concepts that surface discrepancies between social workers’ and the AI’s decision processes [96, 171].

In **Factors the AI did versus did not consider**, we first asked participants, “If the AFST scored a particular case differently than you, would you find it useful to get a heads-up explaining why it did that?” After participants were asked to elaborate on their responses, they were shown a potential version of the AFST that shows examples of factors that were influential in determining the AFST score in a given case, alongside factors the AFST *did not* use but which human workers may have taken into account (see Figure 2.9). In **Factors the worker may not have considered**, we showed participants an interface which highlights factors the algorithm used that the *worker* may not have considered, to support reflection in cases where the

Demographic information	Participant counts
Study participation	Design interviews: All participants except S4 Contextual inquiry: All participants except for S3 and S4 Semi-structured interview: All 13 participants
Study participation dates	First visit-only: C3, C4, C5 Second visit-only: C6, C7, C8, C9, S3, S4 Both visits: C1, C2, S1, S2
Years in current position	Call screeners: <1 year (1), 1-3 years (3), 3-5 years (1), >= 5 years (4) Supervisors: >= 5 years (4)
Have you ever worked in your current position without the AFST?	Call screeners: Yes (4), No (4), Unsure (1) Supervisors: Yes (3), No (1)
How long have you worked with the AFST?	Call screeners: Since the worker's employment (5), Since the AFST's deployment (3), Unsure (1) Supervisors: Since the AFST's deployment (3), Unsure (1)

TABLE 2.3: Participants' self-reported demographics aggregated for call screeners and supervisors.

worker's recommendation is at odds with the AFST score (depicted in Figure 2.10). Whereas the previous two design concepts focus on highlighting potential discrepancies in the *factors* that the worker versus the AFST may have considered in their assessments, in **Type of "risk" predicted by worker vs AI** we invited workers to reflect upon potential discrepancies between their own *prediction targets* versus what the AFST is trained to predict [93]). For example, we asked workers to reflect on whether they resonated with the idea that, based on their personal experiences, the AFST may be predicting a different notion of risk than they are predicting as a worker.

2.19.2 Enabling worker-AI dialogue

In our first visit, participants frequently expressed challenges in integrating AFST risk scores into their decision-making processes, given that they frequently relied on case-specific contextual knowledge that was not accounted for in the AFST risk score [221]. Inspired by these findings, we explored a set of design concepts aimed at making the AFST interface and score calculation process more interactive. We generated two design concepts, inspired by challenges and design ideas that participants had shared during our first visit: **Q&A with the AI**, enabling workers to directly ask the AFST questions about its behavior during use, and **Worker control in AI predictions**, enabling workers to tinker with the set of features that the AFST uses or omits, in order to better understand how certain features are impacting the score.

Factors the AI did versus did not consider

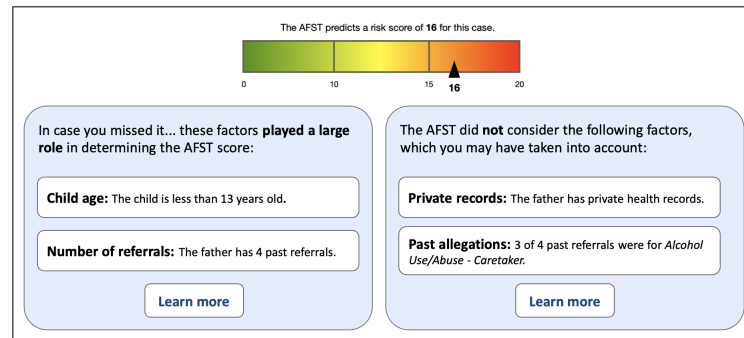


FIGURE 2.9: This design concept invited workers to consider not only factors that contributed to an AFST prediction, but also factors that the AFST is unable to take into account.

Factors the worker may not have considered

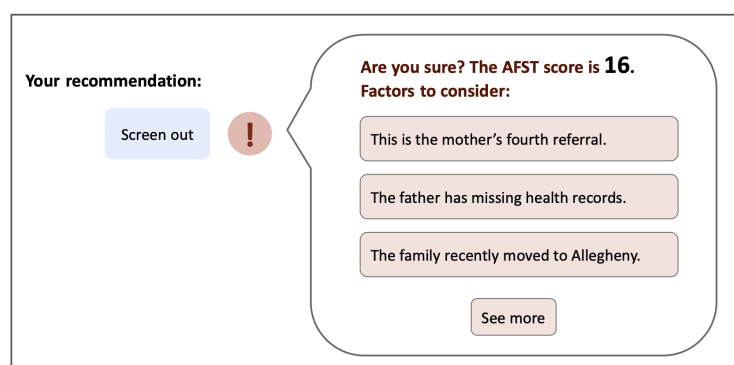


FIGURE 2.10: In this design concept, the worker would be prompted to consider additional factors if their assessment is significantly different from the algorithm's.

Worker control in AI predictions

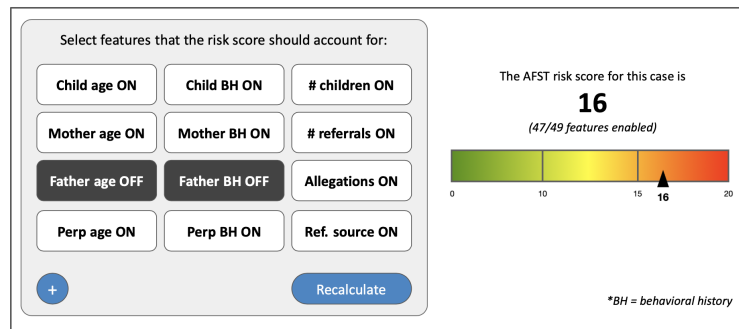


FIGURE 2.11: This design concept, inspired by worker feedback in the first round, would allow workers to dynamically adjust the factors or individuals that the AFST uses when computing a risk score.

While this design topic broadly focuses on revealing what factors the AFST uses to make assessments on specific cases, the two design concepts approach this goal from a different lens. The prompt for the first design concept, **Q&A with the AI**, was intentionally broad: we asked participants to imagine how they might interact with an AFST that was “more human-like,” so it could answer any question they asked. The second design concept, **Worker control in AI predictions** is inspired by findings from the first visit, when discussing with C3 how they might better understand the AFST’s behavior. C3 proposed that it would be helpful if they could more easily explore counterfactuals by “[removing] some things out of the score [... to] kinda play with it and say, ‘if you take this out of there, what kind of score would this person get?’ ” [221]. Therefore, this concept was intended to support workers in exploring the impacts of removing or adding particular features. We first asked participants whether they would be interested in having the ability to control the information the AFST considered to make its assessment, for example, by selecting particular features and removing them from the score calculation, or adding features of importance. Afterwards, they were shown a concrete sketch illustrating one possible way the concept could be implemented (see Figure 2.11), and were asked to imagine how they might use it to inform their screening decisions.

2.19.3 Communicating uncertainty

Prior literature on AI-based decision-support in other high-stakes decision contexts suggests that uncertainty communication (e.g., visualizing the AI’s uncertainty in a particular case) may help experts manage their own uncertainty and better calibrate their reliance on AI recommendations (e.g., [55, 208, 422]. However, much of this literature has focused on tasks where a single, non-contested “ground truth” exists, an assumption that may not necessarily hold in child welfare decision-making [96, 141]. It is thus less clear whether uncertainty communication techniques that have been explored in other domains will transfer well to this this decision-making context.

We presented two design concepts, each exploring a different way for the AFST to communicate its uncertainty to workers in particular cases. **Uncertainty interval** was intended to probe participants’ reactions to a traditional uncertainty visualization, the probability density curve (shown in Figure 2.12). Although many more effective uncertainty visualization techniques have been developed [188, 211], our goal in presenting this design concept was to convey the general idea that the AFST could express uncertainty over a wide range of scores, for a given case. We first

Uncertainty Interval

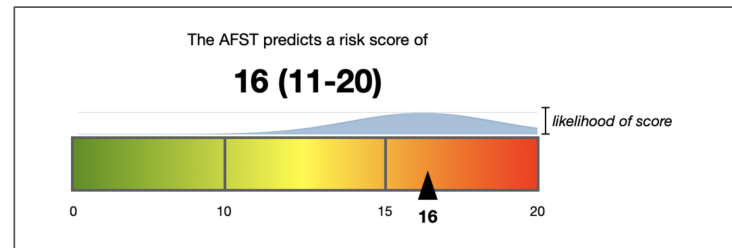


FIGURE 2.12: This design concept draws workers' attention to the inherent variability and uncertainty in outcomes predicted by the AFST, using a conventional probability distribution superimposed on the score.

Uncertainty due to unusualness

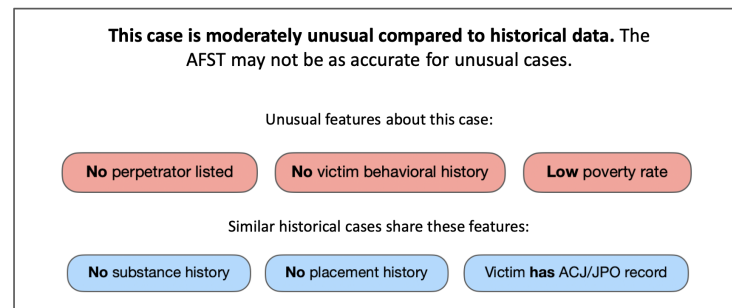


FIGURE 2.13: In this design concept, we explore notions of epistemic uncertainty related to the unusualness of a case, and provide feature explanations of why a case could be considered unusual.

broadly probed participants on whether they were interested in having the AFST tell them “how certain or uncertain it is.” Then, we showed them a concrete sketch of a possible implementation of the concept. **Uncertainty due to unusualness** (in Figure 2.13), on the other hand, considered the “unusualness” of a case, with respect to the data on which the AFST was trained, as a dimension of uncertainty, prompting participants to think about how reliable or unreliable the AFST is likely to be for that particular case [329].

2.19.4 Measurement and feedback on historical decisions

As described in [221], workers described that their existing performance measures caused them to feel organizational pressures to avoid disagreeing with the AFST score “too often,” incentivizing them to rely on the score against their best judgment. While the previous design topics focused on revealing information about the AFST algorithm to help social workers better integrate the AFST into their daily work, this design topic explores ways in which their past decisions with the AFST could be leveraged to improve their use of the tool.

All three design concepts under this topic are intended to explore how participants value receiving feedback about the outcomes of past cases, and what forms of feedback they value. **Similar past case outcomes** shows workers the outcomes of “similar” past cases to help them think through likely outcomes of the current

Similar past case outcomes

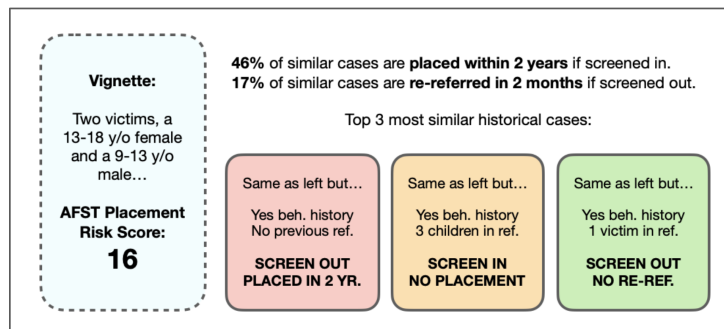


FIGURE 2.14: This design concept portrays AFST scores as founded on patterns in historical data using observed outcomes from similar cases, including the top three ‘nearest neighbors’ of the current case.

Measuring quality of worker-AI decisions

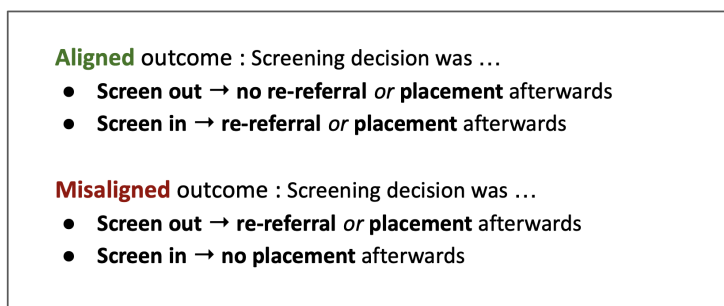


FIGURE 2.15: This design concept maps out a possible way to measure whether a worker’s decision is aligned or misaligned with the final outcome, incorporating commonly-used placement and re-referral metrics.

case they are working on. First, workers were asked whether they would find it useful to see similar historical cases when looking at the AFST score. Then, they were shown one possible visual representation of the concept, as shown in Figure 2.14. By contrast, **Feedback on quality of worker-AI decisions** encourages workers to reflect on their own historical decision patterns, by showing them their recommendations or decisions in actual cases that they themselves processed, alongside the recorded outcomes of those cases. For example, we asked workers to imagine “what kinds of performance or outcome statistics [they] would find meaningful,” if they could see confidential feedback on the quality of their decisions over time. Finally, **Measuring quality of worker-AI decisions** probed participants on what underlying performance measurements and outcomes they would find meaningful, as a foundation for feedback mechanisms like those shown in the other design concepts. Asking specifically about overall decision quality, in addition to individual decision quality, we showed participants one possible measure of human-AI decision-making that is used in existing literature today [82, 93, 96] (see Figure 2.15).

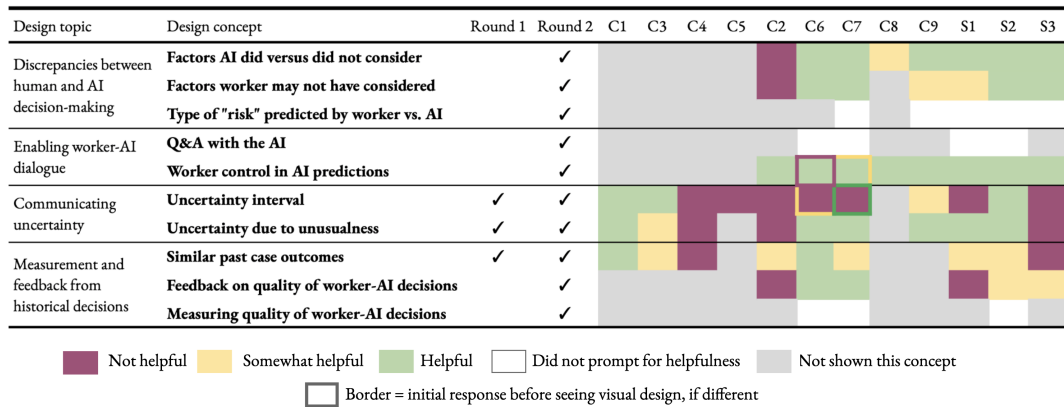


FIGURE 2.16: Matrix showing overall perceived helpfulness for each participant (columns) and design concept (rows). As shown by the checkmarks, three design concepts were introduced during the first visit's design interviews. The remaining seven design concepts were introduced during the second visit's design interviews, informed by findings from the field observations and interviews from the first visit.

2.20 Results

In this section, we present findings organized along our four design topics. Figure 2.16 overviews participants' ratings of overall perceived helpfulness for each design concept. Note that for three design concepts, helpfulness is not shown because the presented concept was intended to support design ideation rather than evaluation; these responses are shown in white. Discrepancies between cell border color and fill color indicate that the participant's assessment of the overall concept (border color) did not match their assessment of the specific design sketch presented (fill color). As described in Section 7.12, because we conducted the design interviews during participants' working hours, some were unable to complete all study activities.

2.20.1 Discrepancies between human and AI decision-making

Workers found it valuable to reflect upon potential reasons for discrepancies between algorithmic judgments and their own, and they wanted ADS interfaces to support them in making sense of such discrepancies. In addition, as discussed below, workers resonated with the idea that the AFST was performing a fundamentally different task than they themselves were performing, which presented a major point of tension in their use of the tool.

Factors the AI did versus did not consider

Six out of eight participants found this design concept helpful, expecting that it could support their understanding of how the score was calculated, enable trust in the score, or better inform their decision to agree or disagree with the AFST score. For example, S2 envisioned that they could leverage their knowledge of relevant context that the AFST lacks, in order to assess the relevance of individual factors that influenced the AFST score. If they had reason to believe that the displayed top factors

were of limited relevance in a particular case, they would then have better evidence to override the AFST's recommendation.

By the time workers run the AFST, they have already assembled relevant information to form their own initial assessment. Given this, C7 suggested that the design should focus on highlighting *"any independent factors, atypical to the normal [factors that workers take into consideration], that have pushed [the score]."* Participants also wanted to supplement their existing mental models of how the AFST works by understanding *how much* the factors impacted the AFST score, not just what factors were used. For example, S1 described a desire to understand how the algorithm weighs factors that may be sources of socio-economic bias: *"... somebody who is getting public assistance, you know, how much weight is given to that [...] versus that person who makes over \$100,000 [...] How much weight is based on income?"* (S1).

Workers expressed interest in using the design concept to identify knowledge gaps between the ADS and themselves. For example, C9 wondered whether the concept could support them in identifying *"missing current concerns for safety"* by showing a list of safety and risk factors to inform their recommendation. Similarly, S2 described that the factors shown could be useful, even if they did not believe they should be considered in the model's risk score. They described that knowing when such factors were considered in the score could help them justify their decisions to disagree with the score.

Factors the worker may not have considered

Overall, participants were motivated to understand why the AFST might disagree with their initial screening assessments. All but one of the seven workers who were shown this design concept found it at least somewhat helpful. Workers appeared open to adjusting their assessments if the AFST could provide reasonable evidence that it was making a better-informed risk assessment than they were. Echoing participants' responses to the previous design concept, **Factors the AI did versus did not consider** to support more informed decisions about when to agree or to disagree with the AFST, participants expressed the need to know how the AFST's algorithmic reasoning diverges from their own, so that they could assess possible shortcomings in both their own and the AFST's assessment.

It was not enough to see a summary of key factors the AFST used: To facilitate such comparison, participants desired the ability to see how the AFST *weighed* various factors, in addition to seeing the set of factors that it used. Furthermore, all seven participants expressed that the usefulness of this kind of interface would depend heavily on the level and kinds of detail it offers to explain particular worker-AI disagreements. For example, one participant argued that the interface would need to allow workers to dive into more concrete details related to each factor, to support workers in personally assessing that factor's relevance to child safety as needed (cf. [50, 175, 264]). After seeing one of the example factors that we presented in our concept sketch, C2 said,

"This is the mother's fourth referral. What is in the report? Because that could be mom's fourth referral on the same incident. These people are sick of her. Like truancy or her kids are playing outside, not doing anything wrong. [...] We get a lot of [...] retaliation reports. Like, 'I call on you. You call on me. I call on you. You calling on me. Calling on you, calling on me.' All of those reports, they keep coming in, keep coming in, keep coming in. And eventually, the [AFST] score is going to be [very high] regardless of what it is."

Similarly, S1 stressed the need to see more detail about each factor: *“if it’s not going to tell me, ‘These are the things that we considered and this is how the score came about,’ [...] then it [does not] mean anything.”*

Type of “risk” predicted by worker vs AI

All five workers who received this prompt resonated with the idea that the AFST is predicting a fundamentally different notion of “risk” than they are. For example, C7 and S3 valued assessing *immediate safety risks* on timescales of days to months, as opposed to trying to assess longer-term risk of child maltreatment on timescales of months to years as the AFST does. C7 raised the point that risk of placement does not necessarily indicate risk of harm, and vice versa, especially in the short term:

“... we don’t really consider, like, when you’re doing a referral, will this child be removed. It’s, like, more like will this child be harmed, or, you know, like, in the immediate... And it’s more safety than risk of removal.”

Similarly, C9 felt that limitations in the AFST’s ability to understand relevant context produced misalignments with their own decision-making goals, and limited the AFST’s usefulness. For example, believing that the AFST predicts risk of system involvement (i.e., risk that a family will use public services), they discussed the algorithm’s inability to distinguish between *positive* versus *negative* engagement with services:

“Cases like for the older teenagers, I mean, there may be at risk to system involvement but, like, the risk or safety are not necessarily as high just because they’re not home as much. [...] It could be the parent also, you know, is involved with mental health or whatever. But that shouldn’t necessarily be a negative, because if they’re, you know, following up with their medical care, that’s not necessarily a bad thing.”

2.20.2 Enabling worker-AI dialogue

Overall, participants responded positively to designs aimed at supporting dialogue between workers and the AFST. Below, we summarize findings connected to these three design concepts, highlighting opportunities to better address workers’ needs.

Q&A with the AI

The four participants who were shown this design concept generated a variety of questions for a hypothetical version of the AFST that could have a “human-like” conversation with them (cf. [264]). Participants envisioned asking the AFST about important factors that were strongly contributing to its risk score calculation. Participants were particularly interested in learning about aspects of a given case that made it atypical or unique among historical cases, in ways that might impact the AFST’s reliability. A couple of participants expressed interest in asking the AFST about trends in its behavior across time and across similar historical cases. For instance, C6 said it would be helpful to see a timeline of previously calculated scores for a case that has multiple past referrals, in order to understand how and why the AFST score evolved over time in response to unfolding information. C7 wanted to see the AFST’s predictions for “similar” historical cases, along with information about the *actual outcomes* of each case. They described how this information could help them better gauge the validity of the AFST score on similar kinds of cases, by

grounding their understanding of the AFST's predictions in subsequent real-world outcomes. C7 also noted that the ability to ask questions regarding the specific data sources the AFST uses could help inform their decision-making in particular cases: *"I would love to know if we could just have a general idea of where the score comes from, just knowing what data sources it's using."*

Worker control in AI predictions

All eight participants who were presented with this design concept expected that it would be helpful in their day-to-day decision-making. Although we presented this concept, in part, as a way to help workers better understand the impacts particular factors have on the AFST score, workers also saw opportunities to use such an interface to imbue the AFST with important contextual knowledge that the AFST might otherwise overlook.

Workers envisioned a wide range of ways they could use this design to help mitigate limitations they perceived in the existing AFST. For example, workers reflected on how the current AFST protocol requires them to include information about individuals who are not relevant to the child's safety, such as collateral relatives or deceased family members who are no longer active in the child's life, when calculating the AFST score. Based on their past experiences using the AFST, they believed these factors disproportionately impacted the AFST score. S1 expressed another interesting goal for omitting certain information from an AFST score calculation: They wanted to mitigate the AFST's use of data from the distant past, for individuals whose present state is no longer well-represented by their past records. In particular, they explained that some families who are currently at low risk for child maltreatment may be unfairly assessed by the AFST based on actions from several years prior: *"[How] can I balance out father went to jail 10 years ago, but now father is out here being a productive citizen doing what it is that he needs to do? But based on where [the AFST is] pulling everything from, it's pulling that, but it's not pulling it that father is now a productive citizen"* (S1).

Two participants (C3 and C7) were initially concerned that they may not be *"educated enough"* (C7) to make such customizations or that they may make biased decisions when omitting variables, especially given their overall skepticism around the current version of the AFST. However, they ultimately expressed interest in the design concept after seeing a concrete design sketch. Both call screeners, along with one supervisor, emphasized that they would like a way to stay accountable for the decisions they make about factors to adjust in the AFST calculation. C3 and C7 suggested that it would be useful to show both the original and the adjusted AFST score to their supervisor in the case report. S2 suggested that the design integrate an open text field where workers must rationalize how they adjusted the AFST score, noting that these accountability measures can both encourage workers to responsibly use the new capabilities offered by the design and, in the process of writing their explanations, give them an additional opportunity to reflect on the validity of their justifications. These ideas align with existing accountability structures at CYF, where supervisors assess call screeners' recommendations when making their final decision and workers regularly have informal, collaborative discussions about case assessments and the AFST score for particular cases.

2.20.3 Communicating uncertainty

Participants' reactions to uncertainty communication was considerably more varied, compared with other design concepts discussed above. Although workers appreciated the idea of raising awareness that algorithmic risk predictions can be uncertain, they had difficulty envisioning how the AFST's uncertainty might impact their own decision-making, given how they currently calibrate their reliance on the tool. As we discuss in more detail below, our findings suggest that in the context of AI-assisted child welfare decision-making, uncertainty communication may be most useful when it maps to a notion of uncertainty that is already familiar and tangible to workers, such as uncertainty due to missing or unreliable information.

Uncertainty interval

Some participants gravitated towards the general idea of communicating uncertainty in the AFST score. For instance, when presented with a sketch of a probability interval over the AFST score, C1 naturally interpreted this in terms of the factual reliability of the input data to the tool:

"Maybe if [past] reports were not founded, you're gonna end up more [on the lower end], but if there's more that were founded or there's more factual information, I can see how it could fall into [the upper end]. I think it gives us more leeway when we're going to make our recommendation."

Some participants expected that visualizations of the AFST's uncertainty in particular cases could help them decide how heavily to weigh the score when it disagreed with their assessment, especially in "edge cases," where they themselves felt highly uncertain (C3, C6). Similarly, C7 and C9 expected that uncertainty visualizations might encourage them to look more closely at the referral documentation to understand a wide score range.

In contrast, other participants expected that an interval-based uncertainty visualization would not affect their decision making, either because they lack trust in the AFST for reasons that such a visualization would not address (C2, C4), or because they would simply ignore the probability representation and use the peak score (S1, S3). In line with this second reason, C5 expressed concern that the probability curve would *"just make me even more confused, I think. I really feel like I'm not all that savvy with this stuff."*

Uncertainty due to unusualness

Participants were more receptive to this formulation of uncertainty than they were to **Uncertainty interval**, given that it aligned with intuitive notions regarding why the AFST might be uncertain, and also because the list of unusual factors could double as a method to explain the score to them. C6 noted that *"when you have a high score but cannot figure out what is possibly generating that – which would probably fit into the unusual referrals – I think for those kinds of things, it could be helpful."* A few participants were particularly interested in the factor "no perpetrator listed" displayed in the concept sketch, since it resonated with their conception of uncertainty as missing information in the referral (C2, C7). C7 recalled a past case in which uncertainty due to missing data led to a disagreement with the AFST score:

"This child had significant behavioral health issues and substantial behavioral health diagnoses... but the services were from another county. So that would've

been missing data... It was one that had to ultimately be screened in even though it was a low risk [AFST score]. So yeah, things that would tell us what's unusual about something would be helpful."

By contrast, S3 was skeptical about highlighting missing information such as "no perpetrator" because supervisors can already assess unusualness by reading the referral:

"If I saw no perpetrator was listed and thought it was unusual, I'd just think it was unusual. I don't need [the tool] to tell me."

Overall, participants were receptive to uncertainty communication approaches that could alert them to data inconsistencies or outliers impacting the score, but only insofar as they believed these approaches could complement (rather than duplicate) their own abilities to detect such features.

2.20.4 Measurement and feedback on historical decisions

Overall, participants had mixed responses to the three design concepts on measurement and feedback. For all three design concepts, some participants expressed discomfort with the ideas presented in the design concept while suggesting ways in which it may be improved; others saw potential in the design concepts' abilities to surface and enable them to benefit from historical case decisions.

Similar past case outcomes

The goal of this design concept was to aid workers' decision-making by allowing them to compare their current case, at decision-time, against similar past cases with known outcomes. Out of the nine participants who were asked about seeing the outcomes of similar past cases at decision-time (**Similar past case outcomes**), six believed that this capability would be at least somewhat useful in informing their decisions, two believed it would not be helpful, and one was uncertain. Participants who saw value in this design sketch believed it could help to standardize decision making across workers (C1, C6), while those who perceived risks in this sketch noted its potential to further perpetuate existing biases in child welfare decision-making [C3, C7, S1, S3]. In the former case, C1 felt that seeing the outcomes of past cases would empower them to feel more confident in decisions that they perceive as particularly high-stakes, because *"you can see trends and statistics and factual information to back up and support what you're trying to recommend."* This desire to justify decisions to others, also observed in some call screeners' and supervisors' responses to **Factors the AI did versus did not consider**, emerged as an important design consideration, given the collaborative nature of social workers' decision-making with the AFST.

At the same time, many participants objected to the underlying assumption that child maltreatment referrals could be compared in the first place. S3 presented a particularly strong critique of the notion of using similar cases to make future decisions:

"Why do I care what's similar? It's not this case. I don't care if your neighbor's case is kind of like yours – it doesn't relate to what's going on at your house... I'm not dealing with your neighbor."

Furthermore, C7 warned that decision-makers might fall into the trap of believing similar cases always lead to similar future outcomes, causing biases to become even further entrenched in their decision processes. Other participants were skeptical

about whether the ways similar cases were retrieved would actually reflect aspects of the cases workers considered meaningful. For example, C2 expressed disbelief in the example metrics shown in the design (which were extracted from real historical cases), explaining that *“it’s just a lot of other variables going on that decide whether or not this child is going to be placed.”* For these participants, the utility of this design would *“depend on how much detail you knew about the previous case”* (C6).

Feedback on quality of worker-AI decisions

In our first visit, we learned that workers currently have little to no opportunities to receive feedback on their AI-assisted decision-making [221]. We used this design concept to explore what forms of feedback supervisors and call screeners would find meaningful, and what impacts they imagined such feedback could have on their decision processes. When asked about the helpfulness of seeing feedback on the quality of their own historical decisions (**Feedback on quality of worker-AI decisions**, four of the six participants presented with this concept expected it would be at least somewhat helpful, and two expected that it would not be helpful. We observed two prevailing narratives: some participants believed the outcomes of cases on which they make decisions are not, and should not be their responsibility; at the same time, other participants saw value in receiving feedback to learn from their mistakes and identify potential biases in their decision making.

Two participants (S1, C2) were not interested in knowing the outcomes of cases on which they make screening decisions. S1 expressed that there are already too many things to worry about in their day-to-day job, so having additional metrics to concern themselves with might not be productive:

“For me, once it leaves the call center, I’m done with it. I don’t care what happens with it. I have no vested interest in those numbers. Once it’s out of here, it’s out of my hands. Because there’s just too many. There’s too much stuff that’s going on currently to have to worry about what’s going on out in another office.”

C2 shared that they occasionally get curious and look up the outcomes for cases they have previously reviewed: *“I keep a list of certain ones that are like, ‘Ooh, what happened?’”* However, they explained that viewing outcomes on past cases does not influence how they make decisions in the future, given the many factors beyond their control that can influence case outcomes. Furthermore, both C2 and S1 expressed fears that any mechanisms to provide feedback to workers on their decision making could readily be appropriated by management and transformed into tools for expanded oversight of their work, which would be used to assess them negatively.

Other participants, including both supervisors and call screeners, expected that receiving aggregate feedback on the quality of their decisions might help them learn from their past mistakes, including by helping them mitigate the impacts of personal biases that may affect their screening decisions. Furthermore, S2 explained that any feedback workers currently receive on their performance tends to be negative, which can decrease workers’ morale over time. As such, S2 saw opportunities for a feedback interface to support greater balance, providing more opportunities for positive feedback on worker decision making.

Call screeners expressed varied goals and desires for feedback on their decision making. For example, C7 was most interested in seeing how well their recommendations aligned with their supervisors’ final case decisions, which are ultimately informed by a call screener’s compiled report and recommendation, as well as the AFST score: *“Usually, I feel like we’re generally on the same page for stuff, but I would like*

to see, like, how our recommendations go versus what, yeah, what is ultimately decided.” C6 said they would be interested in seeing feedback “just for [their] personal growth,” while C7 was excited about using the feedback as a way to encourage light-hearted competition with their peers: “Oh, I’d love [seeing feedback]. We would turn it into, like, a challenge area to see who could be the best worker.” This excitement for gamifying feedback and promoting comparison across workers contrasts with other participants’, particularly supervisors’, concerns that such mechanisms might be appropriated by management to negatively assess their performance.

Measuring quality of worker-AI decisions

This design concept presented participants with one possible method for measuring the alignment between their decisions and actual case outcomes: measuring how often decisions are aligned with re-referral or placement outcomes. Indeed, this measurement approach maps closely onto the AFST’s objective function: the tool is trained to predict the risk of removal and previously predicted risk of re-referral. When presented with this proposed measurement approach, which might be used to support implementation of feedback mechanisms such as those discussed above, three participants ideated possible ways to improve upon this simple performance metric to better capture important nuances in their decision processes. Workers’ ideas regarding how best to measure decision quality aligned with the values that they and others described in response to **Feedback on quality of worker-AI decisions**. As discussed below, they desired measures that would (1) support positive feedback, not just negative feedback (C7), (2) capture the notion that they have greater responsibility for outcomes that occur within a shorter time period, following their decision on a case (S2), and (3) highlight discrepancies between call screeners’ recommendations and supervisors’ decisions (C6).

Workers voiced concerns about naively using placements or referrals to track all case outcomes, regardless of the specific context of a case. For example, C7 characterized placements as usually being the “harshes^t thing [they] do,” and yet placements can also be a positive outcome if the case is severe: “[I would feel good about my past decision] I guess if it’s a severe situation, like, if a child was placed [...] in a family [...] or if a child got services, or if [the case] was closed and then they came back, those are all good [outcomes].” C7 also expressed discomfort with the presented measurement method’s exclusive focus on outcomes that are usually considered negative:

“...if we’re screening something in, we hope to address the issues and avoid re-referral, right? Isn’t that kind of the idea of screening something in [...] I don’t like [using placement outcomes] either. I mean, because if we’re screening it in, we want to avoid placement.”

For these reasons, as C7 emphasized, it may be too simplistic for a measure of decision quality to encode that if a decision is made to screen-in a case, and the resulting investigation yields a placement or re-referral, that this means the choice to screen-in the case was necessarily a “good” decision. Furthermore, S2 raised the point that, over longer time periods, outcomes on placement and referral are increasingly unlikely to be related to their decisions. For example, children might be placed because of a change in family circumstances, such as a single mother passing away, that they have no way of predicting. Thus, this supervisor suggested that a measure of decision quality only considers re-referral and placement outcomes within a 90 day time frame.

2.21 Discussion

In this study, we probed social workers' challenges in integrating a well-established ADS into their day-to-day decision-making. We explored their desires for new kinds of worker-AI interactions using a set of ten design concepts, inspired by workers' prior feedback and design ideas, as well as ideas from prior research literature. These concepts highlighted discrepancies between human and algorithmic decision-making (Section 2.20.1), potential ways to engage in dialogue with an ADS (Section 2.20.2), various notions of uncertainty (Section 2.20.3), and measurement and feedback on historical AI-assisted worker decisions (Section 2.20.4). Our findings suggest fundamental challenges to designing effective ADS in social decision-making contexts, that existing algorithmic tools do not account for. Below, we discuss implications drawn from each of the ten design concepts, followed by a broader reflection on our findings.

2.21.1 Concept Implications

In this section, we discuss implications of our findings, organized by design topic. Throughout this section, we highlight a set of 13 key takeaways (see numbered list items).

Discrepancies between human and AI decision-making

Existing XAI techniques have often focused on communicating model features and weights to support decision-makers in understanding how an ADS arrives at its predictions or recommendations [379]. However, little research has explored the design of interfaces that prompt decision-makers to reflect on factors an ADS *did not* consider, but which may have informed their own decision-making. Through the **Factors the AI did versus did not consider** design concept, we found that social workers were highly receptive to the concept of ADS interfaces that explicitly surface knowledge gaps between the worker and the ADS. An interesting line of future work could be to design ADS interfaces that display explicit prompts about factors that human practitioners commonly take into account in their decision-making, contrasted against the outputs of existing XAI approaches focused on feature importance. The contents of such interface prompts could be informed by in-depth field studies and retrospective data analyses of human decision-making in specific real-world decision-making contexts.

- (1) *Design ADS techniques and interfaces that display an awareness of the factors that a human decision-maker would be expected to use, in contrast to those used by the AI.*

Workers not only viewed the **Factors the worker may not have considered** design concept as helpful for informing their own decision-making, but also for supporting them in explaining and justifying decisions to their supervisors. For instance, if workers believed that certain factors the model took into account were not relevant to a case, or that the model overlooked factors that were critical to a given case, this design concept could support them in justifying their decision to override an algorithmic recommendation. While the idea of re-targeting ADS towards bureaucratic processes has been discussed in the literature [182], the use of AI *explanations* to facilitate collaborative decision-making could be a strategy to improve the usefulness of ADS systems despite retaining imperfect predictive targets.

- (2) *Explore ways for explainable ADS interfaces to support workers' needs in tasks extending beyond decision-making, such as the need to justify agreement or disagreement with the risk score to peers.*

While the above design concepts largely focused on input features, the **Type of risk predicted by worker vs. AI** concept illustrates the tensions that result from ostensibly unjustified misalignments between workers' own decision targets and the ADS's. Future ADS designers should engage workers in co-designing ADS predictive targets and objective functions that workers understand as complementing, rather than clashing with, their existing decision processes. If taking the stance that predictive targets should actually *alter* workers' existing decision processes, as in the case of the AFST, ADS designers should collaborate with workers to jointly identify concrete opportunities to improve existing decision processes and to co-develop a vision for how an ADS can help them achieve these improvements.

- (3) *Engage workers in the co-design of ADS predictive targets that are intended to complement their decision-making.*

Enabling worker-AI dialogue

In line with prior work on human-centered explainable AI (HCXAI) that has asked practitioners what kinds of questions they might like to ask of predictive models, our **Q&A with the AI** design concept yielded a wide range of responses. Many of the questions our participants generated align with common questions identified in prior work, such as the contributions of a given feature in determining AI outputs, the typicality of a given case compared with historical cases represented in the system's training data, or the source of the data that the system learns from and uses [264, 379]. However, workers were also interested in interactions that have been less explored in the HCXAI literature. For example, to inform their decision-making, some workers were interested in *contextualizing the AI model's current output within the history of a given case* by probing how the AI prediction may have changed across time in response to changes in the data for that case. Other workers were interested in gaining a grounded understanding of the AI model's *reliability* on cases similar to one under consideration, a point that is closely related to our design concepts on measurement and feedback.

- (4) *For repeated risk assessments on the same individuals, provide workers with opportunities to understand how the predicted risk may have changed over time.*

The **Worker control in AI predictions** design concept was one of the most favorably-received concepts in our study. While this design concept was intended to help workers understand the impacts of data inputs on the risk score, workers envisioned ways the design could go beyond understanding the score and support meaningful worker discretion and control using their case-specific, contextual knowledge. Prior literature on algorithm-assisted decision-making has discussed tradeoffs between increasing model transparency and opening opportunities to 'game the system.' For instance, Saxena et. al. found that caseworkers tweaked inputs to CANS, a decision support tool, because they felt 'gaming the system' was the only way for them to regain agency over their decision-making process [360]. In our study, however, workers imagined ways to exercise their agency in informing the algorithmic score calculation while still maintaining accountability for their actions. For example, workers proposed expanding their existing call screener-supervisor accountability structure

around this design concept, by requiring that workers provide an open-text explanation for changes made, and by ensuring that both the original and worker-updated risk score are made visible when others in the organization review screening recommendations. It remains an open question what combinations of worker control and accountability mechanisms might support acceptable trade-offs between overcoming algorithmic limitations and neutralizing its potentially complementary effects. Furthermore, the implementation and validation of such user-controllable ADS would be a significant technical endeavor.

- (5) *Investigate the technical feasibility of risk prediction tools in which feature values can be interactively adjusted to account for contextual factors known to the decision-maker.*
- (6) *Design bureaucratic structures that enable the contextual adaptation of ADS predictions while maintaining accountability.*

Communicating uncertainty

Although participants understood that the probability curve was a better reflection of the “fluid nature of the score” [S2] than the current point estimate, most found this sketch unhelpful. These findings largely align with literature on uncertainty visualization showing that decision-makers commonly ignore probability curves when presented alongside point estimates (e.g., [211]). Data visualization research has often focused on this conventional depiction of uncertainty, improving its efficacy through the use of discrete representations and animations [188, 211, 225]. However, our results suggest that seeing a range over possible outcomes may not be the right kind of uncertainty to visualize in this context. By contrast, visualizations that are designed to align with workers’ existing notions of “uncertainty”, such as **Uncertainty due to unusualness** discussed below, may be more actionable.

- (7) *Communicate uncertainty in terms of concrete sources of variation that align with existing decision-maker notions of uncertainty, rather than through generic representations (e.g., probability distributions).*

To our knowledge, unusualness has not been investigated as a form of uncertainty communication in AI-assisted decision making, although it is similar to notions of epistemic uncertainty proposed in machine learning [315]. It is therefore notable that **Uncertainty due to unusualness** resonated with participants more than conventional depictions of uncertainty, both because it tapped into decision-makers’ existing conceptions of uncertainty and because it explained the algorithm’s confidence level using familiar data features. In light of recent mixed findings on the effects of local feature explanations on decision-making performance [bandsal2021does, 472], combining explanations with uncertainty metrics in this way is a promising avenue for future work. Additionally, our discussions with participants around design concepts **Uncertainty interval** and **Uncertainty due to unusualness** highlighted the ways in which workers account for the state of their own knowledge at the time of a decision, in accord with prior work on expert decision-making [55, 208, 422]. Future work should explore strategies to communicate these alternative forms of algorithmic uncertainty (e.g. unusualness and data missingness) to complement workers’ awareness of their own uncertainty.

- (8) *Explore technical methods for not only quantifying ADS uncertainty, but explaining uncertainty in terms of data features.*

- (9) *Design ADS interfaces that help workers make sense of their own uncertainty, and adjust their decision-making accordingly.*

Measurement and feedback on historical decisions

Making inferences based on similarity between cases is a fundamental underlying assumption of statistical risk assessment tools in general, including the AFST. By rendering this underlying assumption more salient to workers, the **Similar past case outcomes** design concept provoked strong reactions of both hope and discomfort—a reflection of participants’ attitudes towards quantitative comparison of families. Building on Cheng et al.’s work using interactive case comparison to elicit fairness notions [64], this design illustrates how greater transparency into an algorithm’s conceptualization of “similarity” can lead to greater confidence that a statistical risk assessment is founded in real data. However, participants’ reactions also highlighted the limits of case comparison—not only because decision-makers believe families are ultimately non-comparable, but also because the axes for comparison are typically limited to shallow, incomplete administrative data.

- (10) *Explore ways to allow workers to (partially) inform risk score calculation by defining contextually appropriate similarity metrics.*

In the concept **Feedback on quality of worker-AI decisions**, we explored workers’ desires for performance feedback on their AI-assisted decision-making. While some workers were excited about the idea of receiving feedback, others expressed disinterest or concern about the prospects of having their decision-making measured. For example, workers in this study expressed fears that any performance measures may be repurposed by their organization for surveillance purposes. This concern surfaced even though we asked workers to imagine that the feedback would only be used for their own benefit and that others would not see it. Moreover, given that workers were already overloaded and overwhelmed, they worried that receiving performance feedback, and having to sift through and interpret it, might simply add an additional burden to their workday. Given that feedback is critical in order to improve decision-making over time, future work should examine how to alleviate these obstacles toward accepting and utilizing feedback.

- (11) *Explore the design of technologies and organizational structures that provide workers with time, space, and incentives to care about and benefit from feedback.*

Prior studies on the impact of ADS tools on human decision-making have often defined performance according to whether the decision taken (e.g., screening in a child maltreatment call for investigation) aligns with an observed outcome (e.g., whether a child is later removed from their home). In our study, when we presented workers with the **Measuring quality of worker-AI decisions** concept, they highlighted shortcomings of this measurement approach, while both validating and nuancing known limitations reported in prior literature.

First, Coston et. al. [82] noted that the action taken may influence the outcome that is observed. For instance, a worker may have been correct in screening-in a family for investigation, despite the absence of future re-referrals, precisely because the resulting investigation connected the family to services that helped to prevent those re-referrals. Workers in our study also expressed concerns with making causal claims about their own decisions leading to observed outcomes, noting that there are many other factors, besides their own decision, that could lead to placement.

Second, in social decision-making contexts, good decision-making is often not synonymous with high predictive accuracy for an observable outcome. In our study, workers also expressed discomfort with associating placement with good decision-making, given the case-specific factors that nuance how an observed outcome should be interpreted. For example, workers described that placement may be interpreted either as the harshest possible outcome, which workers try to avoid, or as a desirable outcome, depending on the child's specific circumstances. Given the importance of observable outcomes in designing effective performance measures, future research should explore how we might balance the ethical and societal tradeoffs between collecting more data and improving measurement quality.

- (12) *Explore improved human-AI decision-making measures that account for both the decision process and outcomes that are reflective of human's decision targets, which may deviate from that of the ADS's objective function.*
- (13) *Explore ways to directly involve workers in co-designing appropriate ways to evaluate their own AI-assisted decision-making processes and outcomes.*

2.21.2 Broader Implications

While the above implications follow directly from the design concepts we devised, participants in our study often engaged with the proposed interfaces at a deeper level, seeing "through" the interfaces to critique the underlying assumptions and objectives of the AFST. In workers' reactions of surprise and distrust as much as in their enthusiasm, we begin to see a complex picture of how frontline decision-makers envision possible partnerships with AI. As we discuss below, these findings address some general challenges in algorithmic decision support, but they are more importantly grounded in the fundamental challenges of *social* decision-making.

For example, participants repeatedly cited their reliance on contextual knowledge obtained through referral histories and conversations, which they used to judge the relevance of particular features of a case. As we observed through the **Factors the worker may not have considered** and **Worker control in AI predictions** concepts, workers were often concerned with the impact of specific individuals on the score, such as a parent who has a long referral history or who does not play an active role in the child's life. In these cases, even the ability to preview how the risk score changes in response to different feature values, a feature that has been explored in prior system-building studies [475], may be insufficient to render the ADS useful. Rather, workers envision the ADS as a way to summarize a risk prediction informed both by administrative data *and* their contextual knowledge, and to communicate these assessments succinctly to collaborators. This kind of future may require ADS algorithms and interfaces that are reconceptualized as communication tools, rather than static predictions. At the same time, new methods of validating ADS and strategies to mitigate 'gaming the system' behavior [360] may help promote accountability and consistency in the use of such systems.

Another recurring topic in discussions with participants was that the metrics used by the ADS to measure decision accuracy do not align with their own, a common characteristic of social decision-making domains [126, 416]. While the predictive target in the current version of the AFST is intentionally designed to nudge workers towards considering longer-term risk (aligned with practices in some health domains [436]), our findings suggest a need to reconsider the validity of the model's

targets in the context of workers' own decision objectives (see **Type of "risk" predicted by worker versus AI** and **Measuring quality of worker-AI decisions**). Future work should explore ways to redesign ADS targets and measures to complement workers' existing decision processes and objectives. Until workers' underlying value misalignments with the model are addressed, interface improvements to the ADS may fail to fully capitalize on humans' and AI's complementary strengths.

Perhaps most fundamentally, workers expressed doubt about core aspects of statistical risk prediction, including the management of uncertainty and comparing families based on administrative data. Workers certainly acknowledged uncertainty in their own decision-making, yet their reactions to the **Uncertainty interval** and **Uncertainty due to unusualness** concepts indicated skepticism that knowing the *algorithm's* uncertainty would help them. While visualizing the uncertainty or confidence of a prediction is often cited as a strategy to calibrate reliance [422, 472], participants in our study preferred to be alerted to uncertainty (potentially beyond the model) that might make the case unusual or uncertain to *them*. For example, surfacing possible data incompleteness issues resonated with workers' notions of uncertainty accumulated during information-gathering. As we observed in the **Similar past case outcomes** concept, workers also questioned the assumption that families' structured referral histories can be compared in valid ways, and maintained that their role was to assess the unique situation of each family. Compounding this finding with observations of workers compensating for algorithmic errors and disparities [61, 96], it may seem that the ADS's role is superfluous in these experts' decision-making process. However, our design concepts offer potential avenues for ADS to provide alternate forms of usefulness, either in its current predictive framing or in one that emphasizes the aspects of information-gathering and collaboration.

2.21.3 Limitations

Our findings reflect the perspectives of workers at a single site, who all utilize the same ADS. However, we note that many aspects of workers' relationship to this ADS may be similar to other AI-augmented social work contexts, particularly given the AFST is often held as a model for how ADS in social work can be implemented. Nevertheless, this group of participants may have notable differences simply because the AFST has been in use for so long. For instance, workers in this context may not have received as much recent training as workers in similar settings due to the algorithm being well-established. It is possible that our design concepts elicited less radical proposals overall, given that workers were quite experienced with the tool as it exists now. On the other hand, it could also be the case that these workers provided more critical feedback and design ideas, given their extensive experience (and daily frustrations) in interacting with the tool.

Additionally, we did not implement interactive prototypes of these design concepts, or evaluate improvements to participants' decision-making if they were to use them. Given workers' concerns about evaluation metrics and the technical challenges of many of our concepts, a quantitative evaluation of decision-making would have limited our ability to test a broad range of concepts. However, a future evaluation of some of our designs at higher fidelity would be crucial to distinguish decision-makers' stated perspectives from their actual uses of ADS.

2.22 Conclusion

By eliciting decision-maker reflections on several design concepts, each reflecting possible ways to augment or redesign an existing ADS, this work provides a new lens into how workers wish to engage with AI in a high-stakes social context. Our findings suggest potential ways that ADS interfaces can better support workers' needs to identify relevant decision factors, manage multiple sources of uncertainty, and collaboratively build justifications for decisions. Overall, our findings demonstrate the continued importance of frontline workers' insights in guiding decision about the kinds of roles new technical systems should or should not take on in complex social decision-making contexts.

Chapter 3

Organization-AI Interaction

This chapter is taken from the following paper:

Anna Kawakami, Amanda Coston, Hoda Heidari, Kenneth Holstein, and Haiyi Zhu. “Studying Up Public Sector AI: How Networks of Power Relations Shape Agency Decisions Around AI Design & Use.” In *Proceedings of the ACM on Human-Computer Interaction* 8, no. CSCW2 (2024).

Abstract. 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* other actors *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, our findings shed light on how infrastructural, legal, and social factors create barriers and disincentives to the involvement of a broader range of stakeholders in decisions about AI design and adoption. Agency decision-makers desired more practical support for stakeholder involvement around public sector AI to help overcome the knowledge and power differentials they perceived between them and other stakeholders (e.g., frontline workers and impacted community members). Building on these findings, we discuss implications for future research and policy around actualizing participatory AI approaches in public sector contexts.

3.1 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 in areas such as social services, public health, criminal justice, and education [23, 261, 439, 467]. 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) [19, 73, 417]. However, the use of AI tools in these domains have been met with significant contention. A growing body of research and public media discusses issues of bias [73, 311], lack of transparency [43, 361, 475], and lack of validity [80] of public sector AI deployments, amongst other concerns. 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 [41, 62, 173, 311], interface-level interventions for frontline workers using these tools [178, 182, 214, 222, 459], and community members' concerns and desires around them [43]. 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 [319, 443]), we still lack an empirical understanding of *how* these decisions are made in practice. Addressing this question requires us to turn our attention to actors with significant power to shape AI design, adoption, and use. 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 [302]. In the study of AI systems, recent work has urged academic researchers to reorient the studies of algorithmic fairness [22] and machine learning datasets [288] 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 [86, 234]). 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.

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. In HCI and CSCW research, much of our knowledge around public sector AI comes from the perspectives and experiences of those directly impacted by higher-up decisions (e.g., frontline workers [221, 360] and community members [43, 395]). To better understand the role of agency decision-makers, we ask the following research questions:

- What factors are agency decision-makers' shape decisions to move forward with a proposed AI project?
- What opportunities exist to improve such decision-making?

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. Our analysis surfaced how *power relations with* and *assumptions about* other actors—situated within, above, and below the agency—shaped their decisions around AI design, adoption, and use. In particular, we contribute insights into:

- **Dynamics and sources of disagreement around AI *within* public sector agencies:** When discussing public sector agencies' decision-making around AI adoption and use, prior academic and grey literature has often treated agencies as monolithic entities. Our findings surface how existing depictions of agencies overlook the intra-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 3.4.1)
- **How pressures from external institutions shape decisions around AI design, adoption, and use:** Our findings suggest that agency leaders and creators of public sector AI tools experience constant pressures from other powerful institutions (e.g., legal systems and state courts, private companies), that shape decisions around AI design, adoption, and use. Whereas prior empirical research has tended to examine the organization as a unit of analysis (e.g., to study organizational governance, worker interaction, or model design), 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 3.4.2)

- **Barriers and disincentives to hearing and empowering impacted communities:** While academic research has proposed a multitude of approaches (e.g., [327, 385]) 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 3.4.3)

By studying up the network of power relations public sector agencies operate within, we advance an empirical understanding of current challenges to actualizing AI development improvements in practice, even when argued 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 and policy communities. In Section 3.5, we discuss opportunities to make this shift through interventions that streamline the adoption and implementation of participatory AI approaches, specifically in public sector contexts. Towards this goal, we discuss how research and policy interventions can reshape—as well as leverage—the network of power relations that public sector agencies reside within.

3.2 Background and Related Work

3.2.1 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 [73], homeless services [247], education [180, 346], and predictive policing [375]. 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 [62, 119], the validity of the models used [80, 336], and the (lack of) support for frontline workers using them [221].

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 [182], 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 [97] (e.g., value-sensitive algorithm design [474], deliberation-based participatory algorithm design [469]) 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. [182] 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., [221, 360]), implicating how design approaches and interventions proposed in the academic literature may, in the future, *actually* fit into real-world contexts [219]. 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 [221]. 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.

3.2.2 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* [22, 127, 288, 369]. 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., [107]).

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*” [302], 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. [22] 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. [288] 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.

3.3 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.

3.3.1 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.

3.3.2 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.

3.3.3 Analysis

The 16 hours of interview recordings were transcribed then qualitatively coded using reflexive thematic analysis [39] by two researchers. We ensured that all interviews were coded by the first author who conducted all the interviews, and, when 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 [29]. 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 (Section 7.13) by these three highest level themes.

3.3.4 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 3.3.1. 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.

3.3.5 Positionality

Our research team collectively holds expertise across human-computer interaction, computer-supported cooperative work, critical computing, public policy, 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 [62, 221, 222] and community members [395], have collectively shaped our perspectives which guide our approach to this research.

3.4 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 [138, 355, 443]. 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 3.4.1), *above* (Section 3.4.2), and *below* (Section 3.4.3) the agency.

3.4.1 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 3.4.1). 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 3.4.1). 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 3.4.1).

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 [355], 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, during the interview, one participant recalled an experience where their agency had developed a new AI tool that outputs risk scores, intended to assist frontline workers' screening decisions around potential child maltreatment cases. The participant, who managed a team of AI developers in their agency, described that this tool was never deployed due to opposition from an 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, during the design activity, 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.

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 who leads their department’s efforts on developing new AI tools 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)

Reflecting on their past experiences talking with frontline workers, another R&D Worker with experience leading the development of new AI tools for their agency described: “It seems that no matter how much we voice that, that concern is present. [It’s] hard [to] have people really grasp that the tool’s meant to help inform their decision [...] 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. The participant, who worked in their agency’s AI development team, described:

“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, during their interview, 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, during the design activity, one frontline worker who was involved in piloting their agency's new AI tool expressed frustration with claims they heard that the AI tool 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.

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 interviews with a couple of agencies, agency decision-makers described that 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. During the design activity, participants described that in current development processes, 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 policies 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.

3.4.2 Power Relations with those Above: Constant Pressures and (Lack of) Support from External Institutions

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 not incentivized to address the agency’s concerns.

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 leading efforts to technologically modernize the region their agency serves 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 [426]) 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, during the design activity, 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.

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)

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)

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). Therefore, the participant advocated for federal requirements to mitigate the risk of private companies simply rerouting their services to other (less ethics-prioritizing) customers.

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. One frontline worker, who piloted an AI tool designed to assist reunification decisions for children placed out of their home, recalled feeling concerned that an attorney would overlook the workers’ judgements in favor of the AI tool’s risk score:

“... 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.

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 expressed during the design activity that they desired 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. For example, one agency leader believed that attorneys were overly “conservative” towards the use of new AI-based tools:

“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)

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. In the next section, we discuss participants’ perceptions of and attitudes towards engaging with other stakeholders to their deployed AI tools—particularly, those situated “below” the agency.

3.4.3 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., [247]). 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 public sector AI development. Overall, we found that individual agency leaders and R&D workers expressed interest in involving community members (Section 3.4.3) 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 3.4.3). 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 3.4.3).

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, during the design activity, 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).

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 during the interview that they received 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. Reflecting on how they want their development process to improve in the future, during the design activity, participants expressed interest in receiving feedback on their AI tools from community members; however, 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, who manages a team responsible for implementing newly developed AI tools into specific decision contexts, 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. 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)

In addition to navigating infrastructural challenges around following appropriate logistical norms (e.g., payment, timing of sessions), participants were additionally concerned about their ability to navigate power and knowledge differentials

between them and impacted communities. We elaborate on these concerns in the next subsection.

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, during the interview, 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). During the design activity, as the participant emphasized the importance of developing AI tools to meet outcomes that center community (rather than agency) needs, the participant expressed further concern 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, as they envisioned potential future challenges during the design activity. 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)

Overall, while participants acknowledged the need to involve impacted community members in their AI development process, in many cases, they felt that their agency was not well-equipped to actualize this need. We discuss implications for addressing these challenges—including the lack of infrastructural and social support in engaging stakeholders—in the Discussion section.

3.5 Discussion

The SIGCHI research community has advanced several methods and tools to support the responsible development of AI tools. One particularly promising form involves expanding participation—from the earliest stages of AI design to after the AI model is deployed—to broaden *whose* perspectives inform AI design and use decisions [97, 243, 452]. In examining how agency decision-makers currently make AI-related decisions, our study strongly suggests that participatory AI methods and tools provide a particularly promising pathway to improving the responsible development of AI tools in public sector contexts. However, the research community’s exploration of these methods and tools in *real-world* organizations, especially in public sector settings, remains nascent. While hurdles to bringing HCI approaches from the research community into real-world practice are expected, identifying ways to streamline this process into public sector contexts remains a relatively under-explored and—as AI tools continue to proliferate across public sector settings [219, 248, 262, 269]—increasingly time-critical challenge.

In situating public sector agencies’ current practices and challenges in a network of power relations, we surface a broader set of considerations—including regulatory, infrastructural, and cultural factors—that play a crucial role in the real-world adoption and implementation of participatory AI approaches. Our findings point to unique implications for how to support participatory AI in *public sector* contexts, and more broadly, suggests opportunities for future research to distinguish between public versus private sector needs and applications around implementing participatory approaches. In the following, we discuss several of these implications (Section 3.5.1) along with limitations and reflections on our research (Section 3.5.2).

3.5.1 Power-Conscious Implications for Designing Public Sector AI

Building on our findings, we propose power-conscious recommendations for researchers and policymakers working towards improving the design and use of public sector AI tools. 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 note that there are both opportunities for future work to better *mitigate* the negative effects of power relations, as well as to *leverage* existing power relations to support more responsible practices. Towards realizing these opportunities, we discuss implications for mitigating and leveraging the network of power relations public sector agencies reside in, to support more participatory AI design, evaluation, and governance approaches. We first discuss research implications to reshape power relations with those “within” and “below”

public sector agencies (Section 3.5.1). Next, we discuss policy implications to leverage existing power relations from “above” to support this reshaping (Section 3.5.1).

Reshaping Power Relations With Those Within and Below

How can the research community better support public sector agencies in expanding participation around AI design, evaluation, and governance? Our findings point to opportunities to help refocus and redirect existing research efforts around public sector AI. In the following, we discuss research and design implications to support public sector agencies in reshaping power dynamics in interactions amongst those within (from agency leaders to frontline workers) and with those below (impacted communities).

- **Make visible the value of participatory approaches to AI design and evaluation to public sector agency leaders and developers**, for example, through case studies that showcase tangible benefits and metrics (whether quantitative or qualitative) that can clearly communicate the value of participatory approaches. This should **increase incentives for agency leaders and developers to explore and uptake participatory approaches**, while also supporting agency workers already doing the work of advocating for increased community and worker engagement (Section 3.4.3). Future research could identify metrics and outcomes that agency leaders find motivating, that additionally demonstrate the benefits of using participatory approaches (e.g., public perception, worker happiness and satisfaction).
- **Ensure agency leaders and developers are aware of existing easily adoptable methods**, by improving the dissemination and findability of participatory methods and tools developed in the research community. Increasing awareness of the vast landscape of existing approaches can help lower agency barriers to exploring their use, by increasing the chances that agency leaders find an approach that meets their context-specific needs. In turn, awareness of the landscape of participatory methods may also help to mitigate misconceptions around *whose expertise* is valuable across stages of the AI development cycle and the practicality of realizing this value. For example, agency leaders were concerned about the feasibility of involving impacted communities without domain knowledge on AI in AI development processes (Section 3.4.3) and believed frontline workers’ involvement should be consolidated to later rather than earlier stages of AI development (Section 3.4.1). In this case, ensuring public sector agencies are aware of approaches for effectively involving non-AI domain experts in early stages of AI design has potential to help mitigate these concerns.
- **Train agency leaders and managers on ways to effectively work with frontline workers and community members**, for example, by creating safe communication spaces for frontline workers and developing the reflexive skills needed to interpret and respond to worker concerns. While prior academic literature and regulatory standards currently focus on training for *frontline workers* using AI tools, our findings suggest the additional need for training interventions at the leadership and managerial levels that could help foster an organizational culture of deliberate and safe cross-role collaboration. For example, the training could teach leadership and managerial workers about different forms of participation (e.g., from the Ladder of Participation [14]),

to help caution against pitfalls like “participation-washing” [383] and scaffold ideation around new policies or procedures that could support meaningful cross-stakeholder participation in their agency. Training agency leaders and managers to create the organizational and social conditions needed for effective collaboration with frontline workers has potential, over time, to help reshape lopsided power relations between agency leaders and frontline workers. This can open more opportunities to fully understand (and avoid misinterpreting) frontline workers’ underlying concerns around AI (Section 3.4.1). Moreover, learning these skills can additionally help address agency leaders’ doubts around their ability to effectively power share with impacted communities during design processes (Section 3.4.3).

- **Innovate on participatory AI methods and tools to better address problems of practice.** Develop methods and toolkits that provide **concrete guidance and support for social and political tasks required**. For instance, these tasks might involve developing a deep, grounded understanding of the historical, social, and political context of the intended AI deployment setting (e.g., to help interpret AI evaluation measures, understand implications of data limitations, identify who is the “stakeholder” of the AI system versus who is a proxy to the stakeholder, amongst other tasks). Responsible AI methods and toolkits today primarily provide guidance on completing technical tasks, implicitly framing the work of AI ethics as one that is mostly technical [453]. Our findings suggest that, if responsible AI methods and toolkits continue to primarily provide support for work that is framed as “technical”, **there is a risk of further reinforcing agency leaders and developers’ misconceptions around the roles and expertise that are well-equipped and necessary for doing responsible AI work** (Section 3.4.1, Section 3.4.1, Section 3.4.3).

Leveraging Power Relations With Those Above

Many of the implications discussed in the prior section can best be realized through additional policy-level interventions. In this section, we discuss policy implications to support the research community and public sector agencies in pursuing the opportunities described in Section 3.5.1.

- **Situate regulatory standards and guidance in public sector agencies’ on-the-ground needs and practices**, for example, by streamlining communication across different levels of government around the personnel and resources currently available versus needed to do responsible AI work. Participants in our study described that their agency currently struggles to adopt “best practices” proposed by the government (e.g., NIST RMF), given adoption requires the creation of new roles and additional resources (Section 3.4.2). While existing federal efforts to support the integration of AI into government services (e.g., AI Factsheet [186]) communicates the importance of providing guidance on AI adoption, our findings emphasize the importance of equipping agencies with the resources needed to follow that guidance. Mitigating resource gaps would require having **mechanisms for public agencies to communicate the (in)feasibility of implementing guidelines and, ideally, to request additional resources or funds**. For example, this can be supported by creating a technology unit in federal government (e.g., similar to the United States Digital Service) dedicated to supporting consultation and collaboration between federal and local agencies. This type of government program can help intermittently

test regulatory standards in collaboration with local agencies, to experimentally implement standards and iteratively learn opportunities for mitigating resource-level challenges.

- **Provide government funding for research priorities that innovate new participatory approaches, including through examining the *integration* of these approaches into specific organizational settings (e.g., public sector agencies).** Few research projects on AI design innovation publicly demonstrate the use and integration of their proposed methods in real-world public sector agency settings. Our findings suggest there are several overlooked problem spaces at the intersection of developing and actually implementing participatory AI design approaches that work at scale across public sector agencies (e.g., around supporting the findability and adoptability of design approaches). Using government funding to incentivize and support research that demonstrates real-world integration and use of participatory AI design approaches can further help lower the barrier for agencies to adopt these approaches in practice.
- **Support public sector agencies in both learning and sharing with others their experiences implementing participatory AI approaches,** for example, through creating a national repository documenting case studies of integrating participatory approaches to AI design into actual public sector agencies. These guides can help public sector agencies more easily learn from each others' experiences around translating research-based design approaches into practice. While participants in our study shared common challenges around developing and procuring AI technologies in each of their agencies, several expressed that they knew little about other agencies' experiences around AI deployment; as a result, they desired knowledge-sharing mechanisms to learn from other agencies' practices (Section 3.4.2).

3.5.2 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 [439]). 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 in their thinking about AI applications and responsible development practices. Therefore, it is possible that the challenges discussed in this paper represent ones that public sector agencies encounter even when they are motivated to prioritize 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 departments (e.g., being under-resourced, having historically tense relations with the community), we believe it is possible that other U.S. public sector agencies may encounter similar barriers and challenges as those described in the paper.

The act of *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 when discussing their responses. However, it is possible that participants were not comfortable being fully transparent about their experiences and challenges. Especially for practices and challenges that involve other stakeholders, a critical line of future work is triangulating observations across a broader range of relevant stakeholders, including frontline workers, impacted communities, and legal experts.

Finally, agency decision-making is often made “public” through online documentation (e.g., [140]) 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, and hope that collaborations between public sector agencies and independent researchers (such as those that made this study possible) become more commonplace.

Chapter 4

Society-AI Interaction

This chapter is taken from the following paper:

Anna Kawakami, Jordan Taylor, Sarah Fox, Haiyi Zhu, and Kenneth Holstein. *AI Failure Loops in the Workplace: The Confluence of Overconfidence in AI and Underconfidence in Worker Expertise*. In *Big Data & Society 2026*. Also in *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*.

Abstract. A growing body of literature has focused on understanding and addressing workplace AI design failures. However, past work has largely overlooked the role of the devaluation of worker expertise in shaping the dynamics of AI development and deployment. In this paper, we examine the case of feminized labor: a class of devalued occupations historically misnomered as “women’s work,” such as social work, K-12 teaching, and home healthcare. Drawing on literature on AI deployments in feminized labor contexts, we conceptualize *AI Failure Loops*: a set of interwoven, socio-technical failure modes that help explain how the systemic devaluation of workers’ expertise negatively impacts, and is impacted by, AI design, evaluation, and governance practices. These failures demonstrate how misjudgments on the automatability of workers’ skills can lead to AI deployments that fail to bring value to workers and, instead, further diminish the visibility of workers’ expertise. We discuss research and design implications for workplace AI, especially for devalued occupations.

4.1 Introduction

Workplace AI systems are often touted by developers and deployers as tools to augment workers’ abilities and improve their work experience [44]. In practice, however, these systems have frequently fallen short of meaningfully supporting workers [4]. In social services, AI-based decision support tools have been introduced for high-stakes decisions, only to be withdrawn following backlash from frontline workers [355]. In healthcare, clinicians ignore AI tools designed to assist in clinical decision-making [113, 257]. In education, AI tutoring systems have added to teachers workloads and compromised their pedagogical autonomy—despite promises to reduce labor and enhance teaching [167, 178, 427]. Across these sectors, AI systems fail to complement the situated, social, and tacit knowledge that underpins workers’ expertise [256].

Prior research in human-computer interaction (HCI) and science and technology studies (STS) offers several explanations for these failures. Scholars have pointed to

issues of poor contextual fit, where AI systems conflict with existing organizational structures and work practices [405, 459]. Others have highlighted flawed assumptions about the nature of work, misaligned values, or inadequate governance, emphasizing how these design shortcomings intersect with broader labor and power dynamics [e.g., 128, 218, 396, 407]. However, one crucial dimension remains under-explored: *occupational devaluation*, the systemic undermining of worker expertise, particularly within professions predominantly occupied by women and people of color [125]. In these professions, expertise is often not fully understood or explicitly recognized by others [187], and devalued both in status and pay even when accounting for education and experience levels [125].

How does the devaluation of worker expertise shape AI design, evaluation, and deployment? Current literature offers little insight into how this dynamic contributes to the repeated failure of workplace AI to genuinely enhance labor. This paper addresses this gap by placing occupational devaluation at the center of our analysis. We conduct a focused review of academic and grey literature on AI deployments across three domains of *feminized labor* in the United States: K-12 education, home healthcare, and social work.

Feminized labor has long been devalued in the U.S. Such work often involves emotional and care-oriented labor [114], is historically labeled as “women’s work,” and continues to be predominantly performed by women and people of color [e.g., 77, 84, 414]. These jobs tend to be low-status within organizations, with limited worker influence over decision-making [83]. Workers in these fields frequently report that their experience is misunderstood by those in more powerful roles [e.g., 465].

We argue that workers engaged in devalued labor are *especially vulnerable to flawed AI deployments*, which often rest on reductive understandings of their work. These systems not only fail to reflect the complexity of frontline labor but can also *further obscure and erode the visibility of workers’ expertise*, creating a negative feedback loop that deepens occupational devaluation. Through a comparative analysis [367] of past scholarship on AI deployments across three feminized labor contexts (social work, home healthcare, K-12 teaching), we articulate *why* and *how* this dynamic emerges.

To do so, we introduce the concept of *AI Failure Loops*: recurring patterns of failure in the design, evaluation, and deployment of workplace AI systems, in which the devaluation of worker expertise leads to ineffective AI deployments—and those failed deployments, in turn, further erode recognition of workers’ skills. These loops are sustained across the AI development lifecycle, fueled by persistent *over*-estimation of what AI can do and *under*-recognition of the complexity and value of worker expertise.

Some of these challenges are familiar: They echo Forsythe’s observations of expert systems nearly three decades ago [128]. When developing expert systems, AI developers excluded social and maintenance activities in their conceptualizations of “work,” leading to AI deployments that failed to reflect experts’ real-world work [128]. Our paper examines how similar challenges have persisted with modern AI, with a focus on the role of *occupational devaluation*.

Today, with recent waves of both AI hype and actual increases in AI capabilities, developers are targeting a rapidly expanding range of socially complex tasks associated with feminized occupations. Drawing lessons from past AI deployments in these contexts, we can expect this will *further accelerate* the gross under-estimations of worker expertise and over-estimations of AI capabilities that has shaped these occupations in the past three decades. This paper responds to calls to redirect this

trajectory toward *pro-worker futures* [e.g., 4], where worker expertise is celebrated within AI practice and society more broadly, and workplace AI systems are designed to truly complement worker capabilities. By centering occupational devaluation in our analysis, we extend these calls, highlighting the need for AI systems that not only avoid harm but actively contribute to the dignity and visibility of labor.

4.2 Addressing Challenges in the Deployment of Workplace Technologies

Researchers in STS, Computer-Supported Cooperative Work (CSCW), and Organizational Science have long studied challenges in the deployment of workplace technologies. Workplace technologies are often designed or introduced by managers in ways that fail to account for the social contexts within which they are deployed [405]. For instance, those in hyper-competitive workplaces [312] may lack incentives to use collaborative technologies [396]. Computing systems can also fail to be used in practice due to workers' limited visibility into systems and the system's limited visibility into workers' contexts [405]. The design of workplace technologies is also highly political—it can prescribe certain ways of working, thus decreasing worker autonomy [407]. In other words, workplace technologies can, in the Foucauldian sense [129], discipline and punish [407].

A growing body of literature has examined how so-called “AI” systems extend these trends and present new challenges [e.g., 167, 360]. As modern AI systems expand the scope of tasks subject to automation, frontline workers find themselves performing new forms of hidden integration labor or “patchwork” [130] to bridge the gaps between what workplace AI systems purport to do and what they are actually able to do [130, 282, 372]. Prior literature has discussed how workers across occupations, from social workers and judges to radiologists, frequently ignore AI-based decision support tools due to inconsistencies with their decision-making practice and challenges around interpretability [221, 257, 331]. These challenges are further exacerbated by flawed evaluations, where workplace AI systems intended to augment or automate complex human decision-making are built upon imperfect ‘ground truth’ labels, leading to over-claims of model performance [256].

AI researchers are increasingly invoking the language of “participation” to support more responsible technology creation [383]. Such calls have roots in Scandinavian efforts to democratize workplace technology development in the late 20th century [298]. These participatory design engagements were facilitated by strong labor protections and partnerships between labor unionists and Marxist computer scientists [391]. However, worker participation in AI development remains rare in practice, particularly within the United States where the socio-political conditions differ from the Nordic model of industrial democracy. Even when stakeholders are involved, workers are rarely empowered to substantially shape the design process, and familiar challenges often recur once technologies are deployed [98, 218].

We build on existing literature on workplace technology challenges, using the *systemic devaluation of worker expertise* as a lens to understand why some AI deployments fail to truly support workers, even in cases where developers intend to design worker-complementary AI.

4.3 Devaluation of Feminized Labor

Notions of what counts as “work” and, relatedly, expertise are socially situated and have evolved across time [397, 408]. These notions are used to justify the societal and economic (de)valuation of human labor. Feminized labor—which often involves care-oriented labor like teaching, home healthcare, or childcare—provides a particularly striking example. Many forms of feminized labor were historically not viewed as “work,” but as acts of love [125]. Even as feminized labor became formalized into paid occupations through social movements, workers struggled to gain legitimacy. Sociologists have theorized several explanations: Employers may justify low pay by conceptualizing feminized labor as “acts of love” provided out of “goodwill” (the “prisoners of love” framework [114]). The association of feminized labor with women may also lead to its devaluation precisely *because* women themselves are devalued in society (the devaluation theory) [114].

Even when accounting for factors like workers’ education level, occupations with a higher percentage of women pay less and have lower social status [125]. The most female-dominated occupations in the United States still tend to involve care-based work, including home care (85%), social work (85%), and teaching at the primary and secondary levels (77%) [77, 84, 414]. Workers in feminized occupations tend to have low status in their workplace, with little say in organizational decisions impacting their day-to-day work—challenges which are exacerbated by a “glass ceiling” on career advancements [83]. These occupations often also face systemic resource shortages and high rates of worker burnout, all amid growing labor demands [114].

In the context of AI in *feminized* occupations, prior work has quantitatively analyzed labor automation trends, and found that many feminized roles—particularly those involving emotional and care-oriented labor—are projected to be among the most “immune” to future automation [e.g., 124, 139]. Yet, these studies leave open *how* day-to-day labor dynamics shape these trends, and what they mean for worker practices and AI. Separately, a growing body of work documents AI failures in specific work contexts (see Section 4.2). Our paper bridges these lines of inquiry, focusing on AI in feminized occupations in the United States. Through case studies across multiple fields of feminized labor, we introduce the concept of *AI Failure Loops* to trace how the systemic devaluation of worker expertise both drives and is intensified by AI practices.

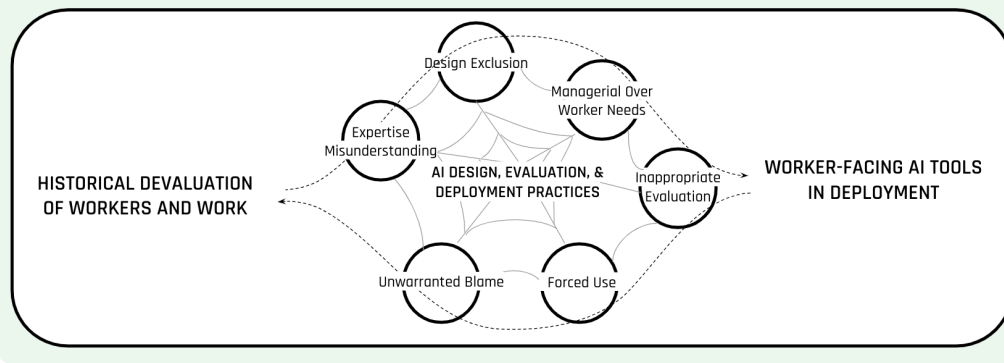
4.4 Methods

To understand how the devaluation of worker expertise shapes AI practices, and vice versa, we conducted a focused review of academic and grey literature on AI deployments across three feminized occupations (social work, home healthcare, K-12 teaching), with analyses informed by our team’s direct empirical experiences. In this section, we elaborate on how this experience informed our case selection (Subsection 4.4.1) and the literature we reviewed in our analysis (Subsection 4.4.2).

4.4.1 Team Experience and Case Selection

We examine case studies of AI deployments across social work, home healthcare, and K-12 teaching, selected because our team has direct experiences collaborating with workers in each occupation. This experience—collectively including over ten years of work observing worker practices, designing interventions with workers, and speaking with leadership and managers to facilitate these collaborations—allowed

AI Failure Loops conceptualizes the negative feedback loop between the *historical devaluation of worker skills and expertise* on one hand, and the *deployment of worker-facing AI tools that fail to bring value to workers* on the other. The six interwoven AI failure modes help make visible the mechanisms by which worker devaluation impacts AI practices, and vice versa.



a

FIGURE 4.1: A visual overview of *AI Failure Loops*. The six failure modes (in circles) that contribute to the dynamics of the *AI Failure Loops* exist within a web to illustrate the inter-connected relationship amongst the failure modes.

us to ground our analysis of relevant literature in our own empirical material (e.g., field notes). For example, our social work case study was informed by our prior findings from field research conducted across five U.S. public sector agencies [218, 221]. This includes experiences observing workers' use of AI-based risk assessment tools [221], speaking with agency leadership and in-house AI developers in social work agencies [218, 219], and designing interventions with social workers [222]. Our healthcare case study was informed by our experiences conducting interviews and design activities with home health aides, aide coordinators, and nurses to understand their perspectives around workplace AI [e.g., 24]. Our K-12 teaching case study was informed by our prior field research studying K-12 teachers' adoption and use of AI-based tutoring systems across hundreds of US-based schools and classrooms [holstein2019, e.g., 178, 179].

4.4.2 Literature Review and Analysis

Our analysis is based on prior literature on AI deployments across the three case studies (social work, home healthcare, and K-12 teaching), each drawn from prior academic and grey literature (including but not limited to prior research from our team). Specifically, as we conducted a comparative case analysis [367] of AI deployments across these occupations, we began to notice common trends, which we explored further through a focused review of academic and grey literature. We gathered and analyzed literature from computing (e.g., HCI, robotics, AIED) and domain-specific (e.g., healthcare, social work, and education) research venues. We searched literature on field deployments of AI systems in workplace contexts [e.g., 205], literature reviews on technologies deployed in specific workplaces [e.g., 244], and syntheses of trends on the impacts of AI on workers [e.g., 167]. Our analysis was not limited to research on challenges from AI deployments. Through our initial comparative case analysis of findings across the authors, and this subsequent review of the literature, we iteratively developed the *AI Failure Loops* concept.

We next present the concept of *AI Failure Loops* (Section 4.5), then describe our three case studies (Section 4.6) to illustrate how *AI Failure Loops* unfold in practice.

4.5 *AI Failure Loops*

AI Failure Loops arise at the confluence of two dynamics: over-confidence in AI capabilities on one hand, and under-confidence in worker capabilities on the other. We conceptualize *AI Failure Loops* as negative feedback loops that occur when the devaluation of worker expertise introduces or reinforces threats to the creation of useful worker-facing AI tools, in turn leading to the deployment of worker-facing AI tools that further amplify harms from worker devaluation. Based on prior literature discussed in the Case Studies (Section 4.6), we identify six interwoven AI failure modes that shed light into how this can happen:

- **Expertise Misunderstanding:** AI tools in devalued occupations are often motivated by misunderstandings of workers' expertise. When designing, evaluating, or deploying AI, this may manifest in workers having their expertise questioned.
- **Managerial Over Worker Needs:** AI tools in devalued occupations are often designed to meet broader organizational goals and targets, at the expense of worker needs.
- **Design Exclusion:** Workers in devalued occupations tend to be excluded from the design process. If they are involved as one of multiple stakeholders in a design process, their perspectives may be considered a second priority to other stakeholders.
- **Inappropriate Evaluation:** Evaluations of AI tools in devalued occupations tend to be designed in ways that advantage AI performance and disadvantage workers, further promoting misunderstandings of their expertise and the nature of their work.
- **Forced Use:** Workers in devalued occupations often do not have the power to refuse technology use.
- **Unwarranted Blame:** Workers in devalued occupations tend to receive unwarranted blame or responsibility for shortcomings or harms observed from AI systems during deployment.

Figure 4.1 provides a high-level visual overview of *AI Failure Loops*. As illustrated by the web, the six failure modes are mutually reinforcing, where the presence of one failure mode can amplify the likelihood of others. As we will see in the Case Studies (Section 4.6), devaluation and misunderstanding of worker expertise are at the core of these dynamics. In our first case, we will show how misunderstandings about how social workers make child maltreatment screening decisions (*Expertise Misunderstanding*) led to the use of AI evaluation measures that were fundamentally misaligned with their decision-making goals (*Inappropriate Evaluation*). Then, we will describe how home health aides are systematically excluded from the design of new technologies (*Design Exclusion*), in part, because they are made invisible by organizational leaders who did not perceive them to be an integral part of the healthcare system (*Expertise Misunderstanding*). Instead, "higher status" healthcare professionals' and institutional needs and goals were prioritized in design efforts (*Managerial Over Worker Needs*). Finally, we will show how K-12 teachers' exclusion from adoption decisions of new AIED technologies (*Forced Use*) disincentivized AI developers from understanding teachers' work practices, limiting their ability to design AI

TABLE 4.1: Additional historical and social context on the three feminized occupations presented in our case studies (social work, home health, and K-12 teaching in the United States), drawn from existing academic and grey literature.

Origins	20th Century	Today
Social Work		
Founded in 1880s by women organizing settlement house movement, a reformist social movement dedicated to bridging resource gaps between rich and poor families [465].	Formalized into profession, with requirements for Master’s training and state licensing to practice in most U.S. states [465].	Social workers continue to be underrepresented among Black or Hispanic [84]. Other fields requiring expertise is misunderstood.
Home Health		
Considered part of domestic life for most families in early 19th century, while professional home care was reserved for wealthy families. By late 19th century, women began organizing efforts to send nurses to homes of poorer families [49].	The Fair Labor Standards Act of 1938 outlaws exploitative practices and instituted an 8-hour workday but home health aides are explicitly excluded [289].	Home health aides continue to be underrepresented among Black or Hispanic [77]. Unlike paid time off, some workers live below team, despite being
K12 Teaching		
Originally conceived as male-dominated occupation that men partook in during farming off-seasons or prior to moving into more well-established professions. Women begin to enter the profession in the late 1880s [352].	Profession is predominantly occupied by women, except among administrators [388]. Profession becomes increasingly formalized with state licensing and university education program requirements [418].	K-12 teachers continue to be underrepresented (e.g., principals, superintendent) and worker devotes similar levels of education.

tools that effectively complement teachers (*Expertise Misunderstanding*). When deployed classroom technologies fall short of achieving their envisioned potential, AI developers and school districts often attribute these shortcomings to teachers’ purported incompetence (*Unwarranted Blame*), with inadequate consideration for potential technology design issues.

4.6 Case Studies

To illustrate the concept of AI Failure Loops, we present three case studies of AI deployments in feminized labor, each emphasizing different failure modes: AI-based risk assessment tools in social work, AI for home healthcare, and AI tutoring systems in K-12 teaching. We provide additional context on each occupation in Table 4.1.

4.6.1 AI-Based Risk Assessment Tools Have Amplified Misconceptions of Social Workers’ Expertise

Context

To address widespread staffing and resource shortages, social work agencies across U.S. state and local governments have started to deploy AI-based risk assessment tools in areas like housing allocation for unhoused individuals and screening for child maltreatment cases. One prominent example of such a tool is the Allegheny Family Screening Tool (AFST), a predictive optimization-based [443] tool intended

to assist social workers in making screening decisions around potential child maltreatment cases. The model is trained on historical administrative datasets to predict the likelihood of a child being mistreated, using the likelihood a child will be removed from their home within two years as an imperfect proxy. Given administrative data on a particular family, the AFST outputs predictions in the form of a risk score between 1 and 20. This class of AI-based risk assessment tool has been adopted across other agencies in the U.S., New Zealand, and Denmark, and applied to other social work decision-making tasks (e.g., foster care placements, temporary lodging placements). In initial years of deployment, the social work agency received both widespread praise and criticism for its deployment of the AFST. On the one hand, the agency was lauded for publishing public-facing evaluations on the ethical implications and technical performance of the model (e.g., [436]). On the other hand, academics, advocates, and journalists heavily criticized the agency for deploying an AI model with biased performance against Black and disabled communities [13, 138] and poor construct validity [80], with some calling to end the use of such tools.

AI Failures in Social Work

We turn our attention to how the kinds of AI design and deployment challenges discussed in these prior critiques may have been influenced by how workers in this domain are perceived and treated by agency leadership and AI developers. This section revisits prior literature on the design and use of the AFST, using *AI Failure Loop* as a lens.

AI tool design fundamentally misunderstands social workers' expertise and overlooks their needs The social work profession has a “glass ceiling,” where workers have minimal opportunities to advance into organizational decision-making roles. As a consequence, there are often large silos between agency leaders and social workers, so those making decisions impacting social workers may have limited procedural understanding of the expertise and on-the-ground knowledge social work requires [183, 221, 360].

In the AFST case, the agency issued an open call to create a new tool to support child maltreatment call screeners in making more efficient decisions. Despite social workers being the main users of the tool, the tool creators did not treat them as stakeholders in the design process. For example, social workers' perspectives on the AI tool were collected only after the model was already developed, using narrowly focused surveys confirming their comprehension of a brief training on the tool, while limiting broader feedback on the tool's overall problem formulation and design [221].

As a result, both researchers and workers have since pointed to fundamental flaws in the AFST's design, which may have been avoided had workers been meaningfully involved in the design process [138, 221]. For example, the longer-term outcomes this AI tool was trained to predict (risk of out-of-home placement over the next two years) fundamentally conflicted with how workers were trained and legally-required to make decisions (i.e., based on assessments of immediate risks to the child). Therefore, once deployed, workers expressed concerns about the tool's validity and uncertainty around how they were expected to interpret and use its outputs [62, 221, 222].

More broadly, workers questioned the validity and usefulness of the tool's overall formulation: statistical risk prediction based on government administrative records.

Instead of presenting opaque predictions derived from the shallow information available in administrative data, workers believed it would have been more helpful to have AI support *them* in making sense of available evidence—leveraging their on-the-ground knowledge and qualitative understanding of case-specific context as human experts [222].

Developers interpret worker complaints around AI as signals of worker deficits

As illustrated above, social workers are often excluded from pre-deployment design and deployment decisions of AI tools that intend to support their work [218]. When workers are given opportunities to provide feedback on AI tools post-deployment, their complaints are often dismissed as an inability to understand a complex technology. For example, in the AFST case, one worker described the types of responses they would receive after submitting feedback forms on the AI tool's risk score as follows:

“The input does not feel like a two-way street. [We] are told why we're wrong, and [that] we just don't actually understand algorithms versus, like, maybe you [the workers] are observing something [our algorithm] missed, you know? [...] In the past, it's felt like when we've taken advantage of that function [to give feedback on the AFST score], it has kind of resulted in just explaining to you why the score is actually right” [221].

Social workers are forced to use AI, even when they disagree Social workers cannot simply refuse technology use [218, 221, 360]. Organizational policies govern how often they can apply discretion with the model. For example, social workers were required to use the AFST and were expected not to disagree “too often.” Workers feared that violating this expectation could jeopardize their employment, so they felt pressure to agree with the AI model, often against their best judgment [221]. Nonetheless, workers still exercised their ability to disagree with the AI on a “budget,” to try to do right by families under these constraints [62].

Similar AI-based risk assessment tools have been deployed in criminal justice. For example, COMPAS assists judges in making recidivism risk assessments [109]. Despite some similarities between the risk assessment tools in social work versus criminal justice (e.g., excluding workers from the design process), we argue there are some notable differences. Given judges' relatively higher-status and organizational power, they have greater ability to refuse technology use by “ghosting the machine” [331], ignoring risk scores at their discretion or discarding the entire risk assessment model, without fear of employment consequences.

4.6.2 “AI for Healthcare” Has Left Behind Devalued Healthcare Experts

Context

Governments and private investors have thrown billions of dollars into healthcare AI innovations—and this number is only increasing each year [117, 459, 463]. Many of these innovations target home healthcare. Yet, today, there exist few, if any, technical innovations that actually *support* home health aides [244]—workers, the majority of whom are women and people of color, who support elderly people and people with disabilities in everyday personal care (e.g., bathing, washing, monitoring health challenges) and household responsibilities (e.g., cooking, cleaning, shopping). Instead, existing technologies in home health have tended to *automate*

worker skills (e.g., care robots [117]), *manage* worker practices (e.g., shift matching [solano2025running]), or provide support to other stakeholders like patients (e.g., interactive voice assistants [24]). For example, to help address the immense care work shortage, several countries—across Europe, North America, and East Asia—have heavily invested in exploring the potential of care robots [117]. In fact, care robots for home healthcare applications are among the main investments in robotics applications [117]. However, a growing body of work has recognized that, if not designed carefully, these systems may disempower workers and amplify challenges they face already [131, 325, 425]. These tendencies to decenter home health aides in AI innovation efforts—in home health and healthcare more broadly—has occurred amid increasing acknowledgement that home health aides are the cornerstone to systems of care, that demands for their labor are increasing, and that institutional support for their wellbeing and labor has remained stagnant [77].

AI Failures in Healthcare

This failure to deploy technical innovations that support home health aides can be understood as a failure of existing efforts to consider the impacts of and on the devaluation of home health aides. This section revisits *why* research and design communities have failed to design AI technologies that genuinely support home health aides, borrowing *AI Failure Loop* as a lens.

Home health aides are rarely designed for In the past decade, “AI for healthcare” has grown substantially as both an area of academic research and capitalist speculation. These efforts often center ways to better meet *doctors’* needs. However, nurses and home health aides—who also play critical roles in healthcare—are rarely designed for or with [244].

While healthcare services rely on collaboration among various professions—including home health aides, nurses, and doctors—most existing efforts to deploy AI technologies overlook this. Instead, the majority of AI for healthcare efforts in the past decades focused on supporting doctors (e.g., physicians, radiologists, surgeons), for example, through new clinical decision support tools [372, 459, 461]. We argue that the decentering of home health aides in “AI for healthcare” efforts has happened, in part, due to misconceptions of the nature of their labor and organizational dynamics that further entrenched these misconceptions.

While home health aides provide many of the same services as nurses in healthcare institutions [162], their labor is performed within their clients’ homes rather than in institutions. Prior literature has described this physical isolation from healthcare institutions has made their labor less visible to other healthcare professionals and policymakers, and created associations with informal home care work [71], challenging the profession’s perceived legitimacy along the way.

These misconceptions of home health aides’ expertise, and along with it, their systemic devaluation, are also observed through how home health aides are treated within healthcare institutions. In *Improving Long-Term Care by Finally Respecting Home-Care Aides* [314], Osterman recalls a conversation with a nurse: “no one pays attention to home-care aides When something changes they notice but are not taken seriously.” These misjudgements may not be explicitly acknowledged nor understood. For example, Osterman recalls conversations with other healthcare professionals and stakeholders, observing:

“other medical professionals, home care agency executives, directors of nursing homes, insurance company executives, and state and federal policy makers, I consistently heard comments that began with a bow to what wonderful people direct care workers are, and then went on to demean their skills and potential” [313]

Even as decision-makers in healthcare (e.g., doctors, hospitals, state regulators, insurance companies) verbally acknowledge home care aides, they “typically do not think that home care aides can be real members of a care team” [313]. Osterman describes this happens, in part, because home care aides’ skills are overlooked, instead misunderstood as that of “unskilled companions [and] glorified babysitters” [313]. Ming et. al. observes that home care aides recognize these misjudgements: “Some people see us as a maid [...] We’re not just here to clean your house [...] Maid are different from an aide” [290]. [290] further cautions that failing to account for the invisibility of aides’ work may contribute to the deployment of AI systems that enforce compliance to impoverished notions of “care”—which, in turn, may further devalue and make invisible home health aides’ work.

In fact, most existing technologies intending to support home care agencies have geared towards supporting higher-status workers in home health agencies (e.g., doctors, nurses), rather than directly helping home health aides. A literature survey from 2022 found only nine studies contributed technologies to support home health aides [244]. Of these, the majority (7/9 studies) were technologies where home health aides were only one of several potential users. The authors found no studies with a long-term evaluation on how the technology impacted home health workers.

Existing technology “for” home health aides have failed to emulate or enhance their work Researchers and policymakers have long looked for ways to sustain the growing population of elderly people—much of which is currently supported through home health aides [27, 161]. While there has been limited research on technologies to support home health aides, there is a comparatively large number of papers exploring ways to automate home-based tasks needed by elderly people, including tasks home health aides usually complete [117]. Many of these technical deployments have come in the form of care robots. However, despite the massive amounts of funding dedicated towards advancing the design of care robots, actual deployments of care robots have often been deemed failures [e.g., 325, 376]. Most of this literature has discussed failures with respect to meeting elderly people’s care needs.

While care robots may have been intended to help alleviate the increasing demand for home care workers, the few studies that account for impacts on care workers have observed impacts in the opposite direction. For example, in a scoping review of 27 papers on the deployment of care robots, researchers concluded the effectiveness of care robots is “mixed” at best. It further observed how robots can instead *increase* caregivers’ workload: care robots “seldom work as a shortcut to increased efficiency or time effectiveness” as workers must undergo the labor of maintaining and operating the robot [325]. Another study found care workers experienced emotional demands, grappled with new ethical challenges, and felt decreased job satisfaction when care robots were deployed in their workplace [455, 456].

Beyond care robots, [solano2025running] discuss similar challenges with other forms of home health technologies, like AI-based shift management technologies that match home health aides with patients. Home health aides reported feeling concerned about how matching algorithms fail to account for the *human* dimensions

of their work, like personality, compassion, and empathy, that impact the quality of matching decisions [solano2025running]. Others described how AI-based shift management technologies may further increase the invisibility of home health aides' work by taking away the already minimal amount of contact they have with home health agencies [solano2025running].

Institutional power imbalances incentivize involvement of higher-status workers in AI design Home health aides are rarely consulted during the design process, even for technologies directly impacting them [244]. This creates a stark contrast with the development of technologies that support doctors. In this section, we briefly discuss how having organizational power as a "high status" worker can create incentives to avoid the *AI Failure Loop*. In doing so, we shed light onto the power and labor dynamics at play in low-status occupations, that lead to deviations in how they are impacted by AI deployments.

The HCI, ML, and medical research and development communities have long been interested in innovating new AI tools to support medical decision-making. Even with decades of attempts, successfully deploying AI tools into real-world healthcare contexts has remained a challenging task [463]. Once deployed, the majority of AI innovations have failed to bring value to doctors [461]. Much like in social work, early versions of decision support tools were often created in isolation from the professional decision-makers themselves: Medical experts were not consulted during the design process. Academics and hospitals found that, even after spending millions of dollars and decades of research innovating on medical decision support tools, doctors were simply ignoring them once they were deployed in practice [459]. This ability to ignore the AI tool—to have full agency over decisions to ignore or disagree with an AI tool—incentivized researchers and developers to consult doctors in future AI innovation attempts.

We argue that doctors' refusal to use AI innovations incentivized researchers and developers to more closely examine *why* doctors were not adopting them [e.g., 458, 459]. To ensure they were creating AI tools doctors would actually use once deployed, the academic and developer communities may have been more incentivized to account for doctors' needs and desires (as evidenced by the gradual surge of academic papers documenting medical experts' needs [e.g., 55, 461] and involvement in AI creation [e.g., 372, 459, 461]).

In comparison, home health aides, like social workers, are not as well-respected in the healthcare system, making it less likely for them to have agency over technology deployments like the ability to refuse technology use. Moreover, because post-deployment evaluations of home health aides' perceptions of AI tools are comparatively scarce [e.g., 244], their negative impacts on workers may also be more likely to go unnoticed.

4.6.3 Persistent Visions of Automation in Education Have Further Devalued Teachers' Expertise

Context

Researchers and developers have long explored ways to introduce new technologies into K-12 education, with hopes to transform the schooling experience. The vision behind AI in education (AIED) research is that every student will be able to work at their own pace and get the personalized support they need [167, 448]. However, these technologies have often been designed with a focus on supporting student

learning, without considering the needs and practical constraints of the *teachers* who are asked to use them in their classrooms [16, 167, 178, 349].

AI Failures in K-12 Teaching

As in the prior two case studies, we see that AI deployments in education have been designed and evaluated at the confluence of two persistent dynamics: policymakers, companies, and school districts have long lauded the potential of AIED technologies to radically transform K-12 education, while, at the same time, these groups have devalued and misunderstood K-12 teachers' expertise. In this section, we examine how AI design, evaluation, and deployment practices have been impacted by these dynamics.

Designing to help students learn, at the expense of helping teachers teach There is a long history of AI development in education guided by the ultimate vision of automating away the role of human teachers [371, 448]. Over the past forty years, research and practice in AIED has largely centered students' needs, often without explicit consideration for teachers' needs and constraints in the classroom [16, 167, 178, 349]. Holmes refers to these deployments as "student-focused AIED"—AI technologies aiming to automate one or more of teachers' tasks for the stated goal of helping students learn, for example, including tutoring systems, virtual writing assistants, automatic writing evaluation, and chatbots. By contrast, Holmes finds that "teacher-focused AIED" has remained largely speculative, confined to academic research prototypes rather than commercial systems [167]. Beyond the research community, commercialization practices help explain the prevalence of student-focused AIED technologies at the expense of teacher-focused AIED. Companies sell AIED technologies to schools and districts, and past research has found that teachers may be given little say in whether or not to adopt a given technology [178, 427].

By reducing K-12 teachers to mere service providers responsible for operating technology and maintaining the classroom [167], many commercialized AIED deployments require teachers to "accommodate" to the technology. Indeed, prior work examining AI deployments observed how teachers often struggle to make AIED technologies work in their classroom, without having to adjust their practices in unreasonable ways or take on new responsibilities along the way [116, 427]. Observing math teachers using an AI tutoring systems in their classrooms, Holstein et al. and Utterberg Moden et al. found that teachers were forced to grapple with numerous contradictions designed into the AIED system, eventually causing them to abandon the system. For instance, while the AI tutoring system aimed to help students work at their own pace, this ran counter to pedagogical practices around collaborative learning and discussion: Teachers had a responsibility to maintain a cohesive classroom environment, where students were roughly synchronized in the content they were learning. [178, 427].

Moreover, these failures to support teachers may be challenging to capture post-deployment because existing AI evaluation measures focus on improving student learning and achievement outcomes with little regard for teacher experiences. As was the case with AI deployments for social workers and home health aides, the emotional and physical labor involved in integrating and operating AI technologies into the workplace are not considered in assessments of new AI deployments.

"Teacher bashing": Teachers receive unwarranted blame for failed AIED deployments Despite the labor and skill required to make AIED technologies work in

their classroom, teachers are disproportionately blamed for limited adoption and use of AIED technologies in classrooms [167, 371, 448]. Indeed, there is often an assumption that the reason AIED tools are not well-implemented is because teachers are too disinterested or unskilled to use the technology effectively. In their book, “Education and Technology: Key Issues and Debates,” Selwyn reflects:

“As is often the case with debates over the ‘failures’ of education systems, ‘blame’ for the restricted use of technology in schools, colleges, and universities has tended to be attributed most readily to the perceived shortcomings of teachers [...] teachers have been deemed to be too old, disinterested, or incompetent to integrate digital technology [...] It is all too easy for enthusiastic academic commentators to indulge in ‘teacher bashing’ and portray teachers as outmoded, obstructive, or ignorant.” [371]

Limited incentive to change AIED development practices Prior work has documented instances in which teachers were able to abandon AIED deployments that failed to bring them value [e.g., 178, 427]. However, even when teachers unite to refuse technology use, they are not guaranteed an escape from *AI Failure Loops* longer term. As discussed in Section 4.6.1 (judges) and Section 4.6.2 (doctors), higher-status workers may be able to simply ignore the outputs of an AI tool, creating natural incentives for future AI developers to account for their needs and desires in future AI creation attempts. On the other hand, lower status workers like teachers may manage to escape from an instance of a failed AI deployment, but these escapes alone are unlikely to prompt meaningful improvements in AI development practices. As discussed above, school districts and developers may still disproportionately blame teachers for low adoption and use of AIED technologies missing opportunities to identify ways to design AI technologies that genuinely support teachers.

4.7 Discussion: Toward Pro-Worker AI

In this section, we discuss implications for researchers and practitioners to support the study and development of pro-worker AI. We begin with three critical questions (Section 4.7.1– 4.7.3) to guide future research, policy, and practice towards a pro-worker future for AI and society. Each question raises implications to help *speak back against* currents that devalue worker expertise through AI practices. Drawing on lessons from our case studies, we explore how *AI Failure Loops* might be avoided by attending to the ways design processes account for power dynamics (Subsection 4.7.1), the ways measurement can be used to make misleading claims about human versus AI performance (Subsection 4.7.2), and opportunities to bridge the gaps between pro-worker AI in theory and in practice (Subsection 4.7.3). We end with a discussion of unique considerations for the case of feminized labor (Section 4.7.4).

4.7.1 (How) can design processes manipulate power?

Our findings demonstrate how the design of workplace AI can reproduce and exacerbate occupational devaluation. The everyday work of devalued occupations, including but not limited to feminized labor, is often obscured from those in comparatively “high status” positions in the same organizations [35]. Yet, across our cases, AI developers routinely privileged the knowledges of “experts” to *speak for*

devalued workers. Rather than involving social workers themselves, developers consulted social work academics (Section 4.6.1); rather than involving teachers, developers consulted school district administrators (Section 4.6.3). In doing so, AI developers further perpetuated the invisibilization of feminized labor and the devaluation of workers' expertise [154]. In failing to include — let alone center — social workers and teachers in the design process, developers created tools that did not work for these frontline workers.

Centering historically devalued workers in the design and research of AI systems poses unique challenges. Echoing longstanding concerns regarding gatekeepers in participatory design [390], relying on managers to access workers may exert undue influence on research and design decisions. Instead of expecting leaders to open a pathway to involve workers, we encourage researchers and practitioners to buy workers' time and form collaborations with existing labor unions and worker advocates [389]. Researchers should also critically examine worker participation in AI development. Workers may feel forced to participate [406] or be included [165] in AI design processes with which they disagree but cannot reject [98]. If devalued workers are involved as one of multiple stakeholders, treating their voice as "equal" to other "high status" stakeholders may reproduce workplace power imbalances [131]. Finally, we caution that managers and developers may recuperate the language of "participation" by using minor worker involvement to justify AI deployments, also known as *participation washing* [383].

4.7.2 (Why) do we need to measure human performance against AI?

The measurements used to evaluate AI models are but one of many representations of the world [37]. Our case studies demonstrate how the design of workplace AI evaluation metrics can further marginalize devalued workers. As we described in Section 4.6.1, managers measured worker versus AI performance according to AI developers' flawed predictive target. Even though this predictive target misaligned with workers' underlying decision-making goals [62], this comparative analysis was used to justify the AI tools' deployment. Moreover, once AI tools were deployed, these flawed measurements were used to track and surveil workers' compliance with AI outputs [221].

Our case studies also strongly suggest that we should question the value of "measuring" workers in the first place. As noted by CSCW scholars, the question of what work to make (in)visible through datafication can be politically fraught [397]. Researchers and developers may find that some work cannot or should not be quantifiable, leading to the development of alternative evaluation standards (e.g., worker satisfaction). However, even in occupations where there has been decades of research and development, as is the case with AIED technologies in K-12 education, the vast majority of evaluations for AI deployments have focused on understanding impacts on other stakeholders (e.g., students), without consideration for frontline workers (e.g., teachers') experiences and perspectives [167].

In contrast to current workplace measurement approaches, future research could explore the design of new approaches *with* workers that help make visible the (currently invisible) impacts of AI deployments on their labor. In these efforts, ensuring worker ownership over data collection and disclosure decisions is critical to ensuring that such measurements can actually enable workers to advocate for better technologies and working conditions (or the abandonment of technologies), rather than further surveil [290].

4.7.3 What happens after worker-centered design?

Our case studies also demonstrate a tension between the theory and practice of pro-worker AI research and design. As illustrated in the education case study (Section 4.6.3), institutions may be less incentivized to adopt or commercialize AI technologies that center devalued workers' needs. This has led teacher-centered AIED tools to remain largely "speculative" [167], where they exist as research prototypes but not as actual tools used in the classroom. Worker-centered AI tools that do get adopted by real-world organizations may still inadvertently reinforce worker devaluation through organizations' governance practices. As *AI Failure Loops* explains, workers in devalued positions tend to receive unwarranted blame for unanticipated issues and harms that arise from AI post-deployment, even if they have minimal power to prevent them. If AI harms are only visible to workers, they may also not have power to refuse to use the technology. Even in efforts to deploy and, when appropriate, commercialize AI technologies, researchers and developers must continue to remain wary of—and actively correct for—the impacts of occupational devaluation.

Researchers can also support pro-worker AI through policy advocacy aimed at empowering workers to govern AI use and maintenance. For example, future work could design new approaches to involve workers in the creation of organizational policies that govern technology use—including to empower worker involvement in decisions around whether and when to decommission a technology. Future work could also explore ways to expose policymakers to the (often overlooked) impacts of technology on workers, to inform the creation of sector-specific worker-protecting regulations (e.g., [468]). Finally, future work should continue to formalize mechanisms for worker collective action by, for example, designing new technologies that allow workers to organize around and share experiences with technology use (e.g., [198]).

Finally, our case studies demonstrate the importance of critical research elevating workers' voices and questioning the design of workplace AI systems. For example, the AFST research we described in Section 4.6.1 partially contributed to some US states stopping the use of similar child welfare screening algorithms in other US states [206]. Likewise, the development of a child welfare risk assessment tool in Denmark was canceled due in part to public scrutiny fueled by critical scholarship [339]. While researchers may not always have the power to ensure that workers are centered in the design of workplace AI systems, they can play a crucial role in contesting the proliferation of workplace AI systems that reinforce worker devaluation.

4.7.4 The Case of Feminized Labor

Redirecting toward a pro-worker future for AI may involve some unique considerations in the context of *feminized labor*. For instance, many feminized occupations center on emotional and care-oriented labor that may be difficult or impossible to quantify, increasing risks of misquantifications in AI development. Moreover, workers in service-oriented positions (i.e., most feminized positions) are often trained to decenter themselves, becoming a "non-person" in their workplace and to employers [128, 290]. Even if AI developers manage to form genuine design partnerships with workers, they may not—without explicit encouragement—advocate for designing technologies that meet their own needs. In past efforts to design technologies *with* feminized workers (e.g., home health aides, K-12 teachers), these workers

often centered clients' needs and de-centered their own needs by default during design engagements, unless nudged by designers to do so [holstein2019, 24]. We anticipate that, with time, redirecting this trajectory will only increase in importance for AI in feminized occupations. When one considers that many of the most devalued occupations are dominated by women, the majority of whom are women of color, it becomes necessary to recognize the presence and prevalence of *AI Failure Loops* as a racial, gender, and social justice issue.

4.8 Conclusion

Visions for “transformative” workplace AI technologies that undermine workers are not new. Technologists have long envisioned that AI technologies would “fix” staff shortages in feminized occupations. The challenges described in this paper—the development of workplace technology without a grounded understanding of workers and their work—are reminiscent of observations made by Forsythe nearly three decades ago [128]. Forsythe found that AI developers building expert systems held impoverished understandings of “work,” disregarding social and maintenance activities. Thus, these expert systems failed to model experts' real-world work practices. Despite the shift from rule-based to machine learning systems, many of Forsythe's observations persist today.

Since Forsythe's writing, the scope and breadth of worker tasks that AI developers target have substantially expanded, from systems attempting to provide emotional support for K-12 students to predicting whether a child is at risk of future abuse or neglect. As this trend continues, our case studies provide a cautionary tale: historically devalued occupations are particularly vulnerable to *AI Failure Loops*. We expect recent increases in both AI hype and actual capabilities will exacerbate the over-estimations of what AI can do and under-estimations of workers' expertise.

Critically, nothing about the future of AI and work is inevitable. To overcome *AI Failure Loops*, we must learn from past developments that perpetuated the devaluation of feminized labor and contest the societal and organizational devaluation of certain work. Re-imagining AI development practices to empower devalued workers may even help foster such cultural shifts. Just as *AI Failure Loops* are self-reinforced through negative feedback loops, reversing these trends may cultivate *positive* feedback loops between pro-worker AI practices and worker valuation. This paper extends discussions of how researchers, policymakers and practitioners can help redirect towards a *pro-worker* future for AI and society, by drawing lessons learned from how worker devaluation has shaped AI practices, and vice versa.

Part II

Chapters 5 - 7: Designing Interventions

Chapter 5

Improving AI Use

This chapter is taken from the following paper:

Anna Kawakami, Luke Guerdan, Yanghuidi Cheng, Kate Glazko, Matthew Lee, Scott Carter, Nikos Arechiga, Haiyi Zhu, and Kenneth Holstein. “Training Towards Critical Use: Learning to Situate AI Predictions Relative to Human Knowledge.” In *Proceedings of the 2023 ACM Collective Intelligence Conference*.

Abstract. A growing body of research has explored how to support humans in making better use of AI-based decision support, including via training and onboarding. Existing research has focused on decision-making tasks where it is possible to evaluate “appropriate reliance” by comparing each decision against a ground truth label that cleanly maps to both the AI’s predictive target and the human decision-maker’s goals. However, this assumption does not hold in many real-world settings where AI tools are deployed today (e.g., social work, criminal justice, and healthcare). In this paper, we introduce a process-oriented notion of appropriate reliance called *critical use* that centers the human’s ability to situate AI predictions against knowledge that is uniquely available to them but unavailable to the AI model. To explore how training can support critical use, we conduct a randomized online experiment in a complex social decision-making setting: child maltreatment screening. We find that, by providing participants with accelerated, low-stakes opportunities to practice AI-assisted decision-making in this setting, novices came to exhibit patterns of disagreement with AI that resemble those of experienced workers. A qualitative examination of participants’ explanations for their AI-assisted decisions revealed that they drew upon qualitative case narratives, to which the AI model did not have access, to learn when (not) to rely on AI predictions. Our findings open new questions for the study and design of training for real-world AI-assisted decision-making.

5.1 Introduction

AI-based decision support (ADS) tools are increasingly deployed to assist front-line professionals in critical, real-world contexts like social services, healthcare, and criminal justice, with the hope of improving decision quality. However, realizing these tools' potential in practice is far from guaranteed [45, 102, 147, 168, 258]. To support human decision-makers in effectively using such tools, it is critical that they receive sufficient training before using them in practice. Yet, today, ADS tools are often introduced into these contexts without adequately preparing the frontline professionals who are asked to use them day-to-day [55, 221, 359].

To address this real-world need, a growing body of studies have explored onboarding and training approaches to support human decision-makers in learning to use ADS tools, demonstrating promising initial results [e.g., 19, 295]. Many of these research studies state a motivation to improve human-AI decision quality in critical applications (e.g., healthcare, criminal justice, and social services). However, a careful review of these studies surfaces key differences between their task designs versus the actual AI-assisted decision-making tasks frontline professionals perform in their day-to-day work [149]:

- **Target-construct mismatch:** Whereas most experimental studies of AI-assisted decision-making assume that human experts and AI models are predicting the same construct, in practice models are often trained to predict an imperfect proxy for the true outcome of interest to human decision-makers [80, 149, 202, 235, 443]. In these settings, effective use of AI predictions depends critically on humans' ability to account for such misalignment.
- **Information asymmetry:** While experimental studies typically assume that humans and AI models have access to the same information for decision-making, in practice humans frequently have access to *complementary* information [160, 171, 172, 221]. In these settings, effective use of AI predictions depends critically on humans' ability to make use of decision-relevant information to which they have unique access.

Although target-construct mismatch and information asymmetry are often discussed in studies of real-world AI-assisted decision-making, their implications for human-AI interaction design remain underexplored. It is an open question how people can best be supported in learning to make AI-assisted decisions in complex, real-world contexts.

In this paper, we explore these questions in a complex, real-world domain involving both properties described above: AI-assisted child maltreatment screening. The use of AI-based decision support tools in child welfare is rapidly spreading [354, 361]. Given significant target-construct mismatch and information asymmetry, it is not advisable in this context to train humans to make decisions that perfectly align with the imperfect proxy used by the AI model [62, 95, 221]. For the same reason, it is unclear how to apply existing measures of "appropriate reliance" on AI predictions in this context, given that these typically assume that alignment with model's target is the goal. To address this gap, we define the notion of *critical use*—that is, humans' ability to situate AI predictions against potentially complementary knowledge uniquely available to them (but not the AI model).

Adopting *critical use* as a lens, we conducted a within-subjects, randomized controlled experiment to investigate *what* and *how* people learn through practice and feedback on decision-making, in the context of AI-assisted child maltreatment screening. We designed training activities that simulate the actual tasks workers face in

a real-world deployment context—Allegheny County’s use of the Allegheny Family Screening Tool (AFST) [74, 119, 221]—but provide participants with accelerated opportunities to practice making AI-assisted decisions, low-stakes setting.

To inform our study design and analysis, we drew upon an extensive body of findings from prior field studies in the AFST context, examining how experienced workers make AI-assisted decisions in this setting [62, 65, 73, 95, 221]. Prior literature indicates that experienced workers in this context use the AFST’s predictions in sophisticated ways, which improve decision-making. For example, quantitative analyses of workers’ decision-making over a multi-year period found that, when workers exercised their discretion to disagree with model predictions on a case-by-case basis, this has the aggregate effect of reducing errors and racial disparities in decision-making, compared with what would have happened had they uncritically agreed with the model’s predictions [62, 95]. In line with our definition of *critical use*, field observations revealed that workers drew upon their knowledge of information unavailable to the AFST (e.g., relevant context from qualitative case narratives) to calibrate their reliance upon the model’s predictions in each case [221]. To support analysis of study participants’ learning, we leveraged our knowledge of experienced workers’ AI-assisted decision-making practices to define *indicators* of critical use in this domain.

Overall, our analyses suggest that our training activities promoted more critical use of AI predictions. In earlier practice opportunities, participants in both conditions started off relying on AI model predictions more heavily. They did so even on cases for which the model erred with respect to the proxy it was trained on. However, with increased practice, participants learned to disagree with the AI predictions more often. In particular, participants learned to make decisions that were more accurate with respect to the decision-making of actual experienced workers. The training also improved participants’ ability to predict the AI model’s behavior. Interestingly, participants who received practice alone saw greater improvements in their ability to predict model behavior, compared to participants who also received explicit feedback alongside practice. Moreover, providing explicit feedback did not impact their learning with respect to decision-making, compared with the effects of repeated practice alone. Our analyses indicate that participants used the qualitative case narratives as a powerful form of *implicit* feedback on the reliability of individual AI predictions, through which participants learned to calibrate their reliance on the AI system. This suggest that, beyond exploring mechanisms for *explicit* feedback on human- and AI-based decisions, future work on training for AI-assisted decision-making should explore the design of rich, implicit feedback mechanisms that empower learners to cross-check AI predictions against concrete, qualitative representations [cf. 394]. More broadly, our findings open up a space of new opportunities and open questions for the design of effective training materials and evaluation metrics for AI-assisted decision-making.

This paper makes the following contributions:

- We introduce the **concept of critical use**, which can be applied to AI-assisted decision-making settings where traditional measures of “appropriate reliance,” based on available ground truth labels, are likely to be unreliable. As a learning goal, critical use is applicable to a range of real-world contexts that involve target-construct mismatch and information asymmetry (e.g., healthcare, criminal justice, content moderation, and education). We explore **how critical use can be measured** via a case study in the child maltreatment screening context.

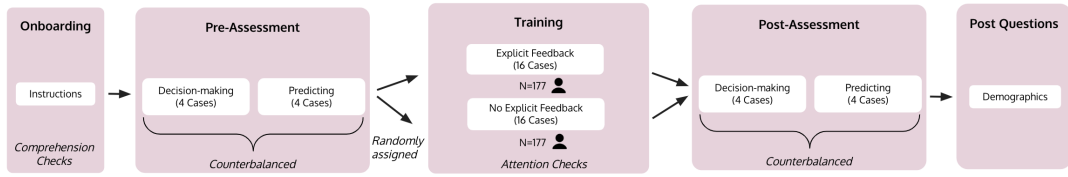


FIGURE 5.1: A high-level overview of the study design. The study included five phases: Onboarding, Pre-Assessment, Training, Post-Assessment, and Post Questions. In the Training phase, participants were randomly assigned between-subjects to the *Practice* or the *Practice + Explicit Feedback* case.

- We present the **first experimental investigation** in the literature of *what* and *how* humans learn through practice and feedback in settings where human knowledge complements that of an AI model (through target-construct mismatch and information asymmetry). Through our analyses, we investigate **how critical use can be supported through training**.
- Based on our findings, we highlight **new opportunities and open questions** for the design of effective training materials and learning measures in real-world AI-assisted decision-making settings.

5.2 Background and Research Questions

In this section, we first briefly overview prior literature on training for AI-assisted decision-making, and then discuss key opportunities to foster human-AI complementarity that are under-explored within this literature. Finally, we motivate and present our research questions.

5.2.1 Training for AI-Assisted Decision-Making

Recent years have seen the rapid adoption of AI-based decision support (ADS) tools to assist human decision-making in settings like social services, education, health-care, and criminal justice [23, 74, 261, 467]. Despite rapid growth in adoption, frontline professionals are often asked to use these tools day-to-day without adequate training. Even in high-stakes contexts, where AI-assisted decisions have the potential to change the trajectory of individuals' lives, the implementation of training approaches has severely lagged behind investments in model development [55, 56, 221, 359]. This lag has also been recognized in recent regulatory and policy efforts, for example, through government directives requiring employee training on AI-based decision-making tools [370].

To address this gap, an emerging line of research in HCI and CSCW has begun to explore the design of onboarding and training interventions to support humans in learning to work effectively with ADS tools. For example, Cai et. al. conducted a qualitative study with pathologists to identify their needs and desires for onboarding to human-AI collaborative decision-making. Other research has begun to investigate the effectiveness of different training approaches via online experiments. For example, Monzannar et. al. introduced an exemplar-based interactive training approach to hone human decision-makers' mental models of a question answering model's strengths and weaknesses, and explored strategies for selecting training

examples that could help humans learn an AI model’s error boundaries most efficiently [295]. Lai et. al. explored model-based tutorials as an intervention to improve humans’ abilities to understand patterns of model behavior, using a case study in deceptive hotel review detections [251]. Overall, this line of work has shown promising early results, suggesting that the proposed training approaches can support human decision-makers in learning to make more accurate AI-assisted decisions.

However, prior studies on training for AI-assisted decision-making have focused on relatively well-defined decision tasks. In the next section, we discuss properties of real-world AI-assisted decision-making that have potential to support human-AI complementarity, but which are not typically represented in experimental study designs.

5.2.2 Potential Sources of Human-AI Complementarity in Decision-Making

While there is an expectation that both human and AI-based judgements are imperfect, there is hope that careful integrations of the two can improve decision-making by drawing upon the *complementary* strengths of each [20, 104, 180, 221, 451, 470]. Prior literature on measurement in AI-assisted decision-making [149, 235, 338] has articulated several properties of real-world decision-making tasks that can support to human-AI complementarity in practice. In this paper we focus closely on two of these::

- **Target-construct mismatch:** In many real-world deployment settings, AI models are trained to predict *imperfect proxies* for the true construct of interest to decision-makers [80, 149, 202, 235, 443]. For example, in healthcare, a widely-deployed ADS tool was trained to predict *medical costs* as a proxy for clinicians’ actual decision-making goal of assessing patients’ *medical need* [297]. However, an analysis found that medical cost is a systematically worse proxy for medical need among certain demographic groups—for instance, hospitals may have historically turned patients away from care due to a lack of insurance or other discriminatory factors [311].
- **Information asymmetry:** In real-world settings, humans and AI models frequently have access to complementary sources of information [171, 221, 235]. For example, clinicians make decisions about medical resource allocation across high-risk patients by drawing upon a patient’s physical presentation, subjective assessments of their well-being, and real-time test results—information that may be unobservable to an AI model [296]. .

Beyond the healthcare examples provided above, these two properties are observed across a wide range of other real-world AI-assisted decision-making settings, including child welfare, criminal justice, online content moderation, education, lending, and hiring, to name a few. Recent field research has shown that, in some settings, frontline professionals are aware of information asymmetries and misalignments between their own goals as human decision-makers versus the proxy outcomes that an AI model is trained to predict. Furthermore, this awareness can shape how they calibrate their reliance on AI predictions on a case-by-case basis [e.g., 62, 168, 221]. Yet to date, most existing experimental research on AI-assisted decision-making has constructed study environments where these properties are absent. In fact, in a recent review of the literature, Guerdan et. al. [149] found that 92% of experimental studies make the assumption that no target-construct mismatch or information asymmetries are present.

Prior studies exploring ways to support humans in learning to make AI-assisted decisions involve tasks such as identifying defective objects [19], passage-based question answering [295], comparing nutritional content [133], or identifying deceptive hotel reviews [251]. In each of these decision tasks, there is no information asymmetry: the human decision-maker is only provided access to the same or a subset of the type of information the AI model is trained on. There is also no pronounced target-construct mismatch: the AI model used in the task is trained on a predictive target (e.g., accuracy of AI-assisted decisions about the *nutritional content* of a given meal [133], or whether an *object is defective* [19]) that directly corresponds to both the AI and the human's predictive target. Like other studies evaluating AI-assisted decision-making [250], accuracy and learning (i.e., changes in accuracy over time) is assessed via correspondence to this ground truth signal.

While prior work indicates that the presence of target-construct mismatch and information asymmetries presents potential for human-AI complementarity in real-world settings, we know little about how to help humans learn to leverage this potential. In this paper, we begin to explore this opportunity space.

5.2.3 Research Questions

Domains involving human decision-making have long recognized that **realistic practice** and **tailored feedback** are key processes for learning. Literature from the learning sciences has demonstrated how practice improves performance across domains, from academic subjects like math and psychology to complex medical procedures and even to professionals' abilities to navigate challenging social interactions [158, 210, 239]. In prior literature, these simple practice effects have been substantially enhanced when supplemented with explicit feedback [111, 238, 239]. In real-world decision-making settings such as social work, healthcare, and education, where there is typically a long lag (on the order of months or even years) between decisions and corresponding outcomes, such simulation-based training approaches have the potential to greatly accelerate learning.

In our study, we investigate *what* and *how* people learn through repeated practice making AI-assisted decisions in the context of child maltreatment screening. In addition, we explore the effects on participants' learning of showing explicit feedback on their decisions. We ask the following research questions:

1. Can **repeated practice** making AI-assisted screening decisions help promote more critical use of AI predictions?
2. Does providing **explicit feedback** on decisions help promote critical use of AI predictions, compared to practice alone?

5.3 Methods

To investigate our research questions, we designed and conducted a randomized online experiment with crowd workers, social work graduate students, and social workers.

Risk Score

The risk score for this case is: **16**



FIGURE 5.2: A screenshot of the AI Risk Score interface shown to participants in the AI-assisted decision-making task. This mirrors the actual AFST interface used by social workers.

5.3.1 Feedback Design

5.3.2 Context

We ground our investigations in the context of child maltreatment screening: a complex, social decision-making context where the use of AI-based decision support tools is rapidly spreading [74, 354, 361]. The task design and data are drawn from a real-world ADS deployment: the Allegheny Family Screening Tool (AFST). As one of the longest deployed and most well-known ADS tools in social services, the AFST has influenced the design and deployment of many other AI-assisted decision-making tools across the U.S. [354]. The Allegheny County Department of Human Services (DHS) deployed the AFST in 2016 to assist hotline call screeners and supervisors in prioritizing alleged child maltreatment cases for investigation [74]. As is common for ADS tools deployed in complex, real-world settings, the AFST predicts readily measurable yet indirect *proxies* for the construct of interest to decision-makers. Whereas frontline decision-makers focus on ensuring children’s immediate safety, the AFST predicts a proxy outcome on a much longer time horizon: a child’s risk of being *placed into foster care* within the next *two years* [62, 221]. The AI-assisted child maltreatment context also involves information asymmetries: prior field studies find that frontline workers’ decisions are informed by rich, qualitative information about a given case (e.g., reported allegations), which is unavailable to the ADS model [221, 363].

In their day-to-day work, call screeners and supervisors at Allegheny County are asked to use the AFST, which outputs a risk score on a scale from 1 (low risk of placement) to 20 (high risk of placement). Our study activities are designed to simulate the actual decision process that call screeners complete. The AI-assisted decision-making task at Allegheny County begins once the call screener receives a call (e.g., from a teacher or neighbor) reporting alleged child maltreatment. The call screener takes notes on the call, and then uses an online data system to collect relevant administrative records on the family being reported. Finally, these administrative records are used to generate the AFST score. Using information from the allegations, administrative records, and the AFST score, the call screener then makes a recommendation about whether or not to screen in the family for investigation. Then, the supervisor makes the final screening decision, informed by the recommendation from the call screener, administrative records, and the AFST score. We describe how our study simulates this task in the next subsections.

5.3.3 Study Design and Materials

To investigate our research questions, we conducted an online experiment with a total of 354 participants, using a pretest-posttest experimental design. An overview of our study design is shown in Figure 5.1. Following a brief onboarding to the study, participants first completed a brief **pre-assessment**, which provided a baseline assessment of their knowledge and decision-making prior to our experimental intervention. Participants then proceeded to a **training phase**, during which they were randomly assigned to practice making AI-assisted decisions either *with* or *without* explicit feedback. Finally, participants completed a **post-assessment**, which was structurally identical to the pre-assessment, but presented participants with new, unique cases. At the end of the post-assessment, participants were asked a small set of post-study questions aimed at understanding their backgrounds. The IQR for study completion time is [48 minutes, 1 hour 25 minutes] with a median of 1 hour 4 minutes.

In the **training phase**, participants practiced making AI-assisted decisions on a series of 16 practice cases, presented in random order. Participants were randomly assigned between-subjects to one of two conditions: *Practice* or *Practice + Explicit Feedback*. As described in the subsections below, for each of the 16 cases shown in the Training phase, participants in the *Practice* condition were asked to review the allegations, administrative records, and the AI model's risk score, and then make a screening decision. Participants in this condition were shown the next case immediately after inputting their decision. By contrast, after making a screening decision, participants in the *Practice + Explicit Feedback* condition were presented with immediate feedback on their decision. After seeing the feedback, participants were then asked to indicate whether they would have made the same decision, "*knowing what you know now*", or whether they would change their decision based on the feedback they saw.

The **pre- and post-assessments** were designed to assess both (a) how participants make child maltreatment screening decisions with AI assistance; and (b) participants' ability to predict what risk score the AI model will assign to a given case. Assessing participants' ability to predict AI outputs, both before and after the training phase enabled us to investigate whether repeated practice and feedback on AI-assisted decision-making would additionally improve their ability to mentally simulate the AI model's behavior [45, 267, 328]. Each assessment included a series of four AI Score Prediction activities and four AI-Assisted Decision-Making activities (which were identical to the decision-making activities presented during the training phase). During the assessment phases of the study, participants were not shown any explicit feedback on either their decisions or their AI score predictions, regardless of which experimental condition they were assigned to for the training phase. In each assessment, we counterbalanced the order in which the two types of assessment activities were shown. For each case shown during the assessment phases, participants were asked to elaborate (via open text) on how they made their AI-assisted decision or AI score prediction immediately after inputting their response.

Below, we describe the designs of the two types of activities included across the training and assessment phases.

AI-Assisted Decision-Making Activity To inform their decisions in the AI-assisted decision-making activity, participants were presented with both a referral vignette and an AI prediction. For each case, the AI prediction and content in the referral vignette was drawn and adapted from case notes on actual cases for which past

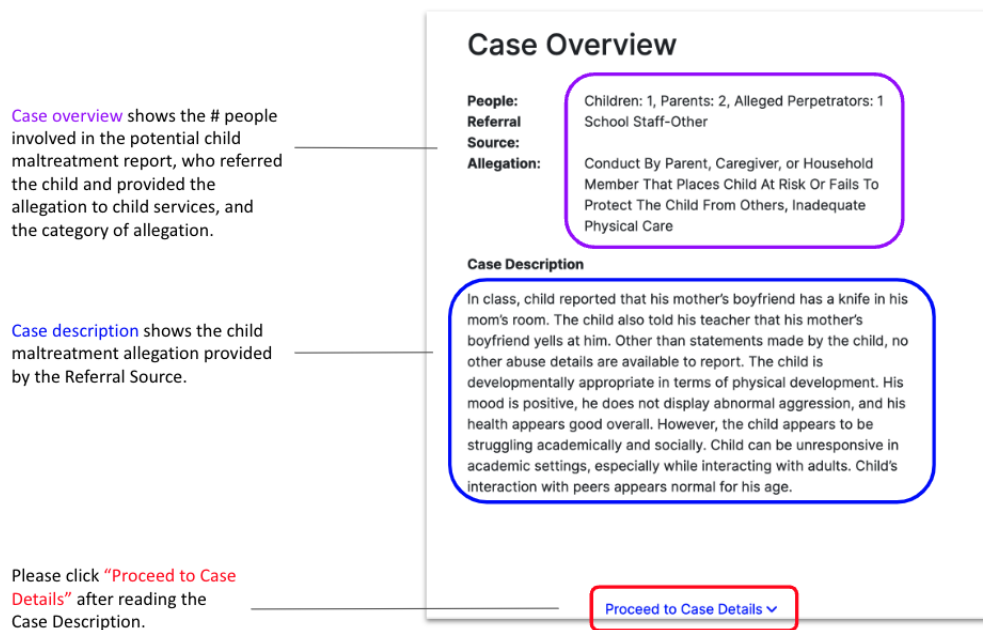


FIGURE 5.3: A screenshot of the Case Overview information panel that provides an overview of the alleged child maltreatment case and a text box with the allegations as a qualitative description. The case content, including the allegations, are taken from real historical data recording past experienced workers' AI-assisted decision-making tasks.

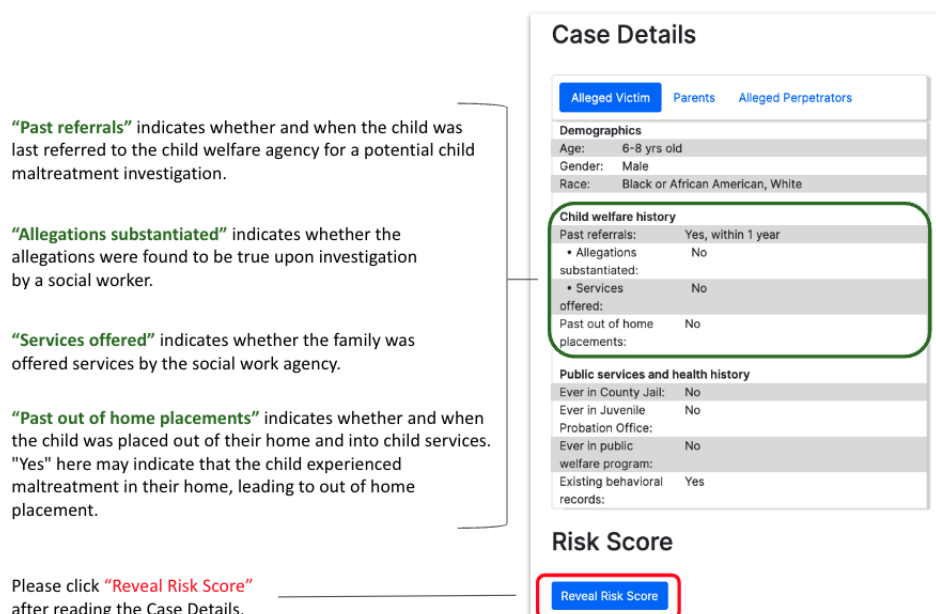


FIGURE 5.4: A screenshot of the Case Details information panel that provides public records and past referral information on the individuals in the alleged child maltreatment case. The content is taken from the same historical case information used to populate the Case Overview and AI Risk Score.

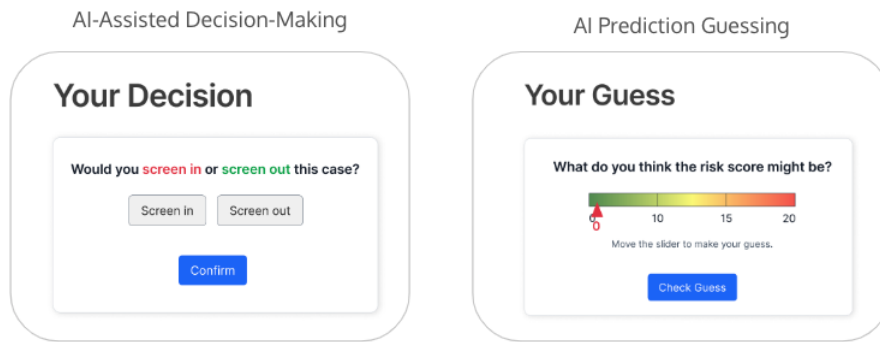


FIGURE 5.5: A screenshot of the question that participants are asked in the AI-assisted decision-making activity (left) and AI score prediction activity (right). To inform this decision, participants also see information from the Case Overview, Case Details, and (for the decision-making task only) the AI prediction.

workers subsequently made AI-assisted decisions, in the Allegheny County context. The referral vignette first shows a **Case Overview** which includes **qualitative allegations**. The child maltreatment allegations were typed by experienced workers based on calls they received (e.g., from a neighbor or a teacher). For each allegation drawn from the historical dataset, we removed any potentially identifying information (e.g., details pertaining to specific regions or households), while still preserving high-level properties of the allegations, as this information could be relevant to the interpretation of other case-related information, including the AI prediction. Figure 5.3 shows a screenshot of a **Case Overview** section for a sample referral vignette. Next, we showed **Case Details** (Figure 5.4) including demographic information, child welfare history, and public services and health history of the alleged child victim, parent(s), and alleged perpetrator(s). For the purposes of our experiment, we opted to show a smaller subset of demographic information than actual workers are able to access in practice, both to exclude potentially identifying information and to ensure that participants would have sufficient time to look through the information (cf. [63]). Finally, we showed the **Risk Score** (Figure 5.2): the actual AFST score for the given historical case.

The presentation order of the Case Overview, Case Details, and Risk Score was designed to reflect the actual order in which social workers encounter each information source in the AFST deployment context, when reviewing a case. We designed the training interface to gradually reveal this information, requiring participants to click ahead in order to reveal the next information source.

AI Score Prediction Activity In the AI Score Prediction activity, participants were similarly shown referral vignettes (including the Case Overview and Case Details) that replicated the decision-making tasks of actual social workers using an ADS. However, rather than showing the AI Risk Score, participants were asked to make their own prediction on the AI output on a scale of 1 to 20. Figure 5.5 (right) includes a screenshot of the question participants were asked after seeing the referral vignette.

Participants in the *Practice + Explicit Feedback* condition received immediate feedback on their decisions during the study's training phase. Given that we do not typically have access to a single, reliable "ground truth" label in the context of child

maltreatment screening, as discussed above, we explore the learning effects of providing participants with multiple feedback signals in two primary forms: *observed outcomes* and *historical human decisions* (Figure 5.6). By showing multiple, imperfect “ground truths,” we provide the human decision-maker with opportunities to reflect on disagreements amongst the ground truths and their own initial decisions, to learn how to more appropriately calibrate reliance on the ADS model over time.

Feedback in the form of *observed outcomes* that were available in the historical administrative dataset, included the following. For each, we briefly summarize its limitations as a feedback signal:

1. Whether the child was **placed within two years**: In the AFST context and in our study, this is the AI model’s predictive target. If the child was eventually placed out of their home within two years, this may signal that the child was maltreated and should have been screened in and subsequently investigated. However, there are other unobservable factors unrelated to the screening decision that may have led to this outcome [80]. For example, the child’s parents may suddenly pass away, requiring the child to be placed into foster care in the absence of any reliable grandparents or relatives. Thus, placement in two years is an imperfect proxy for the accuracy of the screening decision.
2. Whether the child was **re-referred to the agency within two years**: Similar to (a), a range of other unobservable factors unrelated to the decision may have led to this outcome. For example, certain stakeholders (e.g., divorced parents doing “retaliation calls”) may feel incentivized to call the hotline and alleged maltreatment, when in fact, the child is safe [221].
3. Whether the **allegations were substantiated upon screen in**: Only cases that were historically screened in can observe this outcome. Allegations of potential child maltreatment being substantiated based upon investigation may indicate that the child was indeed maltreated, and thus screening in was the “right” decision. *But*, allegations, even if substantiated, may result from cultural misunderstandings or implicit biases from the caller [62].
4. Whether **services were offered to the family upon screen in**: Only cases that were historically screened in can observe this outcome. A family that is offered services may indicate that the child was maltreated, and thus was offered services that can support recovery. In this case, a decision to screen in would be “right.” *On the other hand*, even if a child was not maltreated, a family may receive services as a preventive measure to minimize risk of future maltreatment.

Feedback in the form of *historical human decisions* included:

1. Whether the **call screener’s recommendation** was a screen in or screen out: This is the call screening recommendation a past experienced social worker made, based on the case information and AI risk score they saw.
2. Whether the **call screening supervisor’s final decision** was to screen in or screen out: This is the final decision that a past experienced social worker made, based upon a review of case information, the AI risk score, and the call screener’s recommendation.

Taken together, this set of feedback signals captures both the accuracy of the AI model’s predictions (with respect to the proxy outcome it is trained to predict), as

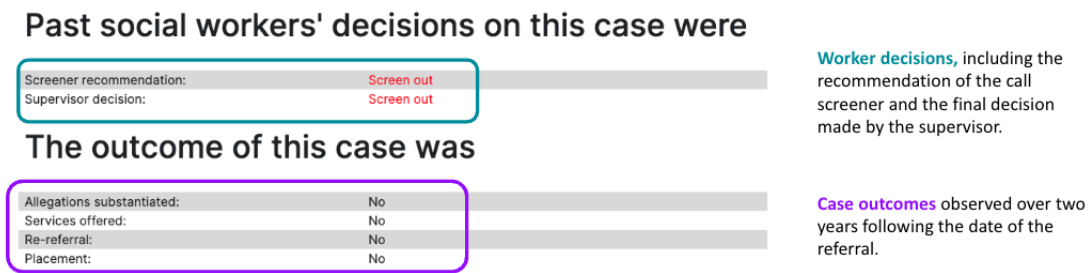


FIGURE 5.6: A screenshot of the six imperfect proxies (screener recommendation, supervisor decision, allegation substantiation, services offered, re-referral, and placement) shown as feedback to participants in the AI-assisted decision-making task in the *Practice + Explicit Feedback* condition. Participants in the *Practice* condition were shown the next case immediately after submitting their decision.

well as what past experienced human workers believed was the right decision after reviewing both the case information and the AI prediction.

5.3.4 Case Selection

The cases shown in the training and assessment activities are informed by the historical administrative records of referrals at Allegheny County between June and December of 2019. In this section, we describe how cases sampled and distributed across phases of the study.

Case Categorization

We developed four high-level case categories (and eight sub-categories) to support the systematic sampling of historical cases for use in our study. Our taxonomy for selecting cases consists of the following alignment-measures:

- **Worker-AFST alignment:** Cases where the worker agreed with the AFST decision¹ (i.e., both state screen in or screen out).
- **Worker-outcome alignment:** Cases where the worker's assessment aligns with the observed outcome (i.e., screen in & placed or screen out & not placed).
- **AFST-outcome alignment:** Cases where the AFST's assessment aligns with the observed outcome.

To guide the sampling of cases, we then constructed a 2x2 confusion matrix breaking down the three alignment measures into four categories as shown in Figure 5.7.

Case Sampling and Randomization

We randomly sampled cases from historical referral data according to our case categorization. We drew 6 cases from each of the 8 sub-categories, yielding a total of

¹We designate cases with an AFST score ≥ 15 as "screen in" and cases with an AFST score < 15 as "screen out". This screen-in threshold aligns with the high risk cutoff shown to call screeners on the AFST tool interface [436]. Use of this threshold also follows prior analyses of automated AFST decisions [62]

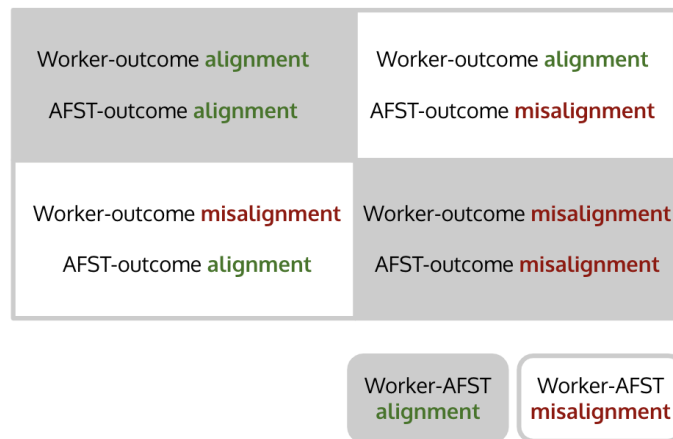


FIGURE 5.7: A 2x2 confusion matrix breaking down the three alignment measures into four high-level categories.

48 cases. Cases were sampled with equal weighting across categories to enable us to study learning effects relative to historical worker and algorithm behavior. Cases within each category were then randomly assigned across the pre-assessment, training, and post-assessment phases of the study. Across participants, cases were assigned to the same study phase (i.e., pre-assessment, training, or post-assessment) to enable item-level analyses, but were presented in randomized order within each phase.

5.3.5 Participants and Recruitment

We recruited a total of 354 participants through the crowdsourcing platform Prolific, social media, and direct email. To explore potential impacts of domain expertise, we sought to include participants with domain knowledge in social work in our participant pool. First, we reached out via email to social work professors at US-based universities and directors of US-based social work non-profits, and asked them to share our study invitation with social work graduate students and/or practicing social workers in their networks. Second, we advertised the study on relevant social work and child welfare groups such as “Social Work Network” and “Child Welfare, Child Protection and SACWIS/CCWIS Technology Professionals.” Finally, we used Prolific to recruit both participants with and without social work backgrounds. To recruit Prolific participants with background in social work, we restricted participation for the study to workers who had completed an undergraduate or graduate (MA/MSc/MPhil) degree in social work. On Prolific, we further restricted participation to workers with a minimum approval rate of 90%. Participants were compensated \$13 for completing the study. Participants who also had social work domain expertise were compensated an additional \$7 for a total of \$20.

All participants were based in the U.S. and were above 18 years of age. Of our 354 participants, 103 participants had domain expertise in social work (i.e., the participant indicated they are currently or were previously a social work graduate student or a practicing social worker). Importantly, only 5 of these 103 participants with social work domain expertise indicated that they have expertise in child welfare specifically, and only 12 participants with social work domain expertise indicated they had prior experience using an AI tool to assist their decisions. However, overall, the 103 participants that have domain expertise in social work may still have more relevant

domain knowledge for the training tasks than the general population of workers we recruited from Prolific. We further discuss implications of our recruitment criteria on the interpretation of our findings in the Discussion (Section 5.5.3).

The study included two comprehension checks and two attention checks that were used to exclude participants from the final dataset. The design of the comprehension and attention checks followed Prolific’s policies². Participants who failed to answer at least one of the comprehension checks were navigated to the end of the study. Participants who failed at least one of the attention checks were excluded from the final dataset.

5.3.6 Measurement and Analysis

Drawing upon findings from prior field studies [e.g., 62, 221], we defined a set of *indicators* of critical use in the context of AI-assisted child maltreatment screening, overviewed below.

Decision alignment measures To quantitatively study changes in participants’ decision-making, we defined two decision alignment measures, covering both the “human” and the “AI” portion of the past recorded AI-assisted decisions used in our study. Because our experiment uses real, historical data (as described in Section 5.3.4), this enables us to explore how participants’ decisions compare to past decisions made by workers experienced in AFST-assisted decision-making. The first of our measures captures agreement with the *AI model*, while the other captures agreement with *experienced workers*:

- **Model Agreement:** We say there are *increases* in Model Agreement if participants come to agree with the AI prediction more often.
- **Worker-based Accuracy:** We say there are *increases* in Worker-based Accuracy if participants’ decisions come to agree with past experienced workers’ decisions more often.

Our main analysis examines participants’ learning between the pre- and post-assessments with respect to each of these two measures. In particular, we examined interactions between the assessment phase (Z_{phase} , a binary indicator with pretest = 0 and posttest = 1) and both participants’ prior domain knowledge (Z_{prior}) and the presence of explicit feedback during training ($Z_{feedback}$). We also included the participant ID e_{pid} as a random effect, to account for the nesting of responses within participants:

$$\hat{Y}_m \sim 1 + Z_{phase} * (Z_{prior} + Z_{feedback}) + e_{pid} \quad (5.1)$$

We supplemented our pre-post analyses with analyses of participants’ learning across the entire study. In particular, we fit mixed effects models to examine interactions between the practice opportunity number Z_{num} (ranging from 1 - 24) and the presence of explicit feedback $Z_{feedback}$ (a binary indicator). To analyze differences across case categories, we additionally examined interactions between the case type and the practice opportunity number. As described in Section 5.3.4, each case category applied to 50% of the 24 total cases, making it possible to run these analyses on balanced amounts of data.

²<https://researcher-help.prolific.co/hc/en-gb/articles/360009223553-Prolific-s-Attention-and-Comprehension-Check-Policy>

Process-oriented measures To study *process-oriented* aspects of critical use, in addition to analyzing our decision alignment measures we used Equation 5.1 to examine pre-post changes in participants' ability to mentally simulate and predict what the AI score would be in particular cases. Participants' mean squared error on the AI Score Prediction activity is shown as "Guessing Error" in Table 5.1). We also qualitatively analyzed (1) participants' open text explanations for how they made their decisions in the pre- and post-assessments, and (2) their self-reported decision-making goals, collected at the end of the study.

Standard accuracy measure: Proxy-based accuracy

The standard measure that is used in existing research literature to measure appropriate reliance is accuracy with respect to the *model's predictive target* [149], which we call Proxy-based Accuracy. We say there are *increases* in **Proxy-based Accuracy** if participants' decisions become more accurate with respect to the model's predictive target. For a given case, a decision is *accurate* with respect to the model's predictive target if the participant either screened in a case that resulted in out-of-home placement in two years or screened out a case that did not result in out-of-home placement in two years.

5.3.7 Positionality

Collectively, the authors hold research expertise across human-computer interaction, learning sciences, statistics, machine learning, computer-supported cooperative work, and critical computing studies. The lead author and co-senior authors have research experience observing and designing interventions with child welfare workers who use AI-assisted decision-making tools in their daily work. The study design in this paper reflects both knowledge from published research literature and the authors' knowledge gained through field studies. None of the authors are affiliated with a child welfare agency; all research was conducted independent of any particular child welfare agency.

Importantly, we explore training towards *critical use* in this paper because we believe doing so is crucial in many settings where AI systems are deployed. As we have argued above, AI performance in complex, real-world settings is bound to be imperfect, and some limitations are to be expected—including some degree of target-construct mismatch and limitations in the information a model is able to access or interpret. Training towards critical use may mitigate downstream negative impacts of AI model limitations. However, we advise such training only in settings where the deployment of an AI-based decision support tool is actually justifiable in the first place. Training alone *cannot* overcome fundamental design flaws in the AI model or surrounding social systems. While we examine AI-assisted child maltreatment screening as an example of a highly complex social decision-making setting in this study, we note that existing AI deployments in this domain have seen many fundamental design critiques, including prior research by members of our team [62, 80, 119, 221, 222, 363, 443].

5.4 Results

Overall, we found that participants in both conditions started off relying on AI model predictions more heavily in earlier practice opportunities. Early on, participants aligned their decisions with the AI model prediction even on cases for which

	Indicators of Critical Use			Standard Metric
	Model 1 Model Agreement	Model 2 Worker-Based Accuracy	Model 3 Guessing Error	Model 4 Proxy-Based Accuracy
Fixed Effects				
(Intercept)	0.65*** (0.02)	0.50*** (0.02)	6.13*** (0.17)	0.73*** (0.02)
Phase (post vs pre)	-0.41*** (0.03)	0.08** (0.03)	-3.03*** (0.22)	-0.16*** (0.03)
Prior knowledge	0.00 (0.03)	-0.03 (0.03)	0.08 (0.25)	-0.02 (0.03)
Feedback	0.04 (0.02)	-0.03 (0.03)	-0.26 (0.22)	-0.03 (0.03)
Prior knowledge × Phase	0.02 (0.04)	0.02 (0.04)	0.42 (0.31)	0.01 (0.04)
Feedback × Phase	0.03 (0.03)	0.07 (0.04)	0.61* (0.28)	0.05 (0.04)
Observations	2832	2832	2832	2832
Marginal R^2	0.153	0.019	0.017	0.105

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

TABLE 5.1: Coefficient estimates for models examining pre-post learning effects. Each cell shows the coefficient estimate followed by the standard error for a given term (rows) and model (columns). A positive *Phase* coefficient indicates an increase in the measure in the post-assessment as compared to the pre-assessment baseline. A visual summary of these findings is reported in Table 5.2.

the model failed to accurately predict the proxy on which it was trained. However, with more practice, participants in both conditions were more likely to make decisions that disagreed with the AI prediction.

Strikingly, we found that participants learned to disagree with AI predictions more often, in ways that aligned more with past frontline workers' decisions, even in the *Practice* condition, where participants did not receive any explicit feedback about past frontline workers' decisions. An examination of participants' open-text feedback suggests that this occurred because participants learned through practice to integrate case-specific, qualitative details from the allegations when making their decisions, with a focus on assessing nearer-term risks to the child's safety (in contrast to the AI model's focus on predicting a proxy outcome, placement, on a longer time window). Past research suggests that this mirrors the way experienced workers make decisions in the field [62, 221]. In turn, participants' decisions began to diverge from model predictions and resembled those of past experienced workers, with increased practice (Section 5.4.4).

Through repeated practice making AI-assisted decisions during the training phase, participants in both conditions learned to more accurately predict how the AI model would behave on particular cases. Compared to participants who received both practice and explicit feedback on their decisions, participants who received practice alone saw more improvement in their ability to predict the AI model's behavior.

Moreover, receiving explicit feedback on decisions did not impact participants' learning with respect to *decision-making*, compared to the effects of repeated practice alone. Our analyses suggest that the qualitative case narratives present in both conditions are a rich information source through which participants can learn to calibrate their reliance on AI predictions. Thus, even in the absence of *explicit* feedback

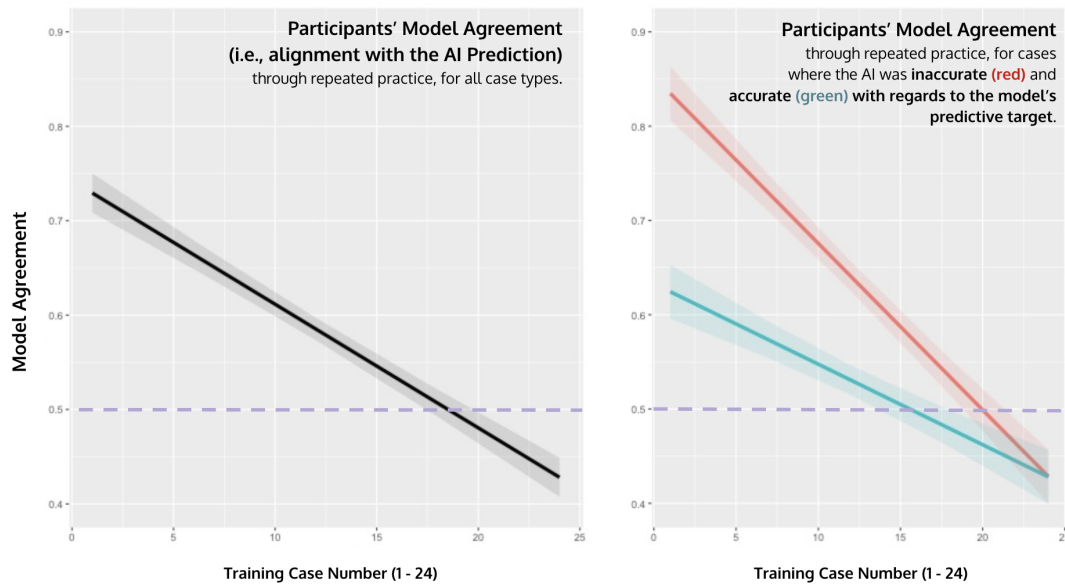


FIGURE 5.8: Mixed effects regression visualizing participants' agreement with the AI prediction, across the 24 training cases. Shaded regions indicate standard error. The purple dotted line indicates average accuracy of the AI model with respect to its (proxy) predictive target.

on decisions: a training interface that provides accelerated practice, with opportunities to cross-check AI predictions against qualitative narratives, appears to be a stronger baseline than expected.

In the following subsections, we first present findings from quantitative analyses of participants' learning and decision-making (Sections 4.1, 4.2, 4.3, 4.5). In Section 4.4, we present complementary findings from a qualitative examination of participants' decision-making *processes* across practice opportunities. Table 5.1 includes aggregate outcome-based findings, and Table 5.2 provides a summary of the directionality of outcome-based findings across the two conditions.

5.4.1 Model Agreement: Participants learn to disagree with AI predictions more often

- (Finding-1) **Participants disagree with the AI prediction more after repeated practice.**
- (Finding-2) **No effect of explicit feedback on participants' agreement with AI predictions.**

Participants started the study by relying on the AI prediction more often, including in cases where the past experienced worker disagreed with the AI prediction. But over time, participants *were less likely to get nudged by the AI prediction*. Instead, over time, participants began to disagree with the AI prediction more often for all case types. As shown in Figure 5.8, in the first case, over 70% of the participants agreed with the AI prediction. However, in the 24th case, only 42% of the participants agreed with the AI prediction. This finding was corroborated by the learning effects regression (Table 5.1), which showed a significant decrease in Model Agreement between pre- and post-assessments as indicated by the study phase coefficient (Coef.=-0.407; $p < .001$; 95% CI: [-.46, -.35]).

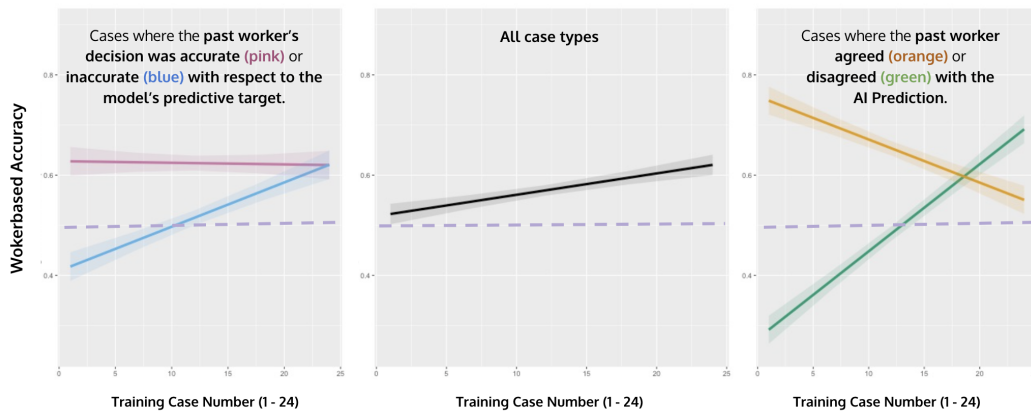


FIGURE 5.9: Mixed effects regression visualizing participants' learning across repeated practice, with respect to the worker-based accuracy measures (i.e., accuracy with respect to past experienced workers).

5.4.2 Worker-based Accuracy: Participants become *more* aligned with experienced workers

- (Finding-3) **Participants' decisions become more accurate with respect to experienced workers' decisions after repeated practice.**
- (Finding-4) **No effect of explicit feedback on the accuracy of participants' decisions with respect to experienced workers' decisions.**

With practice, participants' decisions began to align more with the final decisions of past experienced workers. Overall, Figure 5.9 (middle) shows that, at the start of the training, 52% of the participants made decisions that were accurate with respect to the past experienced worker. By the end of the training, 61% of the participants made decisions that were accurate with respect to the past experienced worker. This finding is consistent with learning effects regression results, which show a significant increase in Worker-based Accuracy scores between pre- and post-assessments (Coef.=0.083; $p<.01$; 95% CI: [0.027, 0.140]).

With increased practice, participants' decisions were especially likely to be accurate with respect to the past experienced worker when *disagreeing* with the AI prediction. As shown in Figure 5.9, for the category of cases in which the past experienced worker disagreed with the AI prediction (right), for the first case, only 30% of participants made decisions that were accurate with respect to the past experienced worker and in disagreement with the AI prediction. By the 24th case, nearly 70% of participants made decisions that were accurate with respect to the past experienced worker and in disagreement with the AI prediction. On the other hand, for the category of cases in which the past experienced worker agreed with the AI prediction, at the start of the training, nearly 75% of participants made decisions that were accurate with respect to the past experienced worker. By the 24th case, this proportion went down to just over 50% of participants.

Additionally, participants' decisions became more accurate with respect to the past experienced worker regardless of whether those past experienced workers made decisions that were inaccurate or accurate with respect to the model's predictive target, an accuracy-based measure that is widely used in prior studies evaluating the accuracy of AI-assisted decision-making with this AI model, the AFST

	Indicators of Critical Use			Standard Metric
	Model 1 Model Agreement	Model 2 Worker-Based Accuracy	Model 3 Guessing Error	Model 4 Proxy-Based Accuracy
<i>Practice Condition</i>	↓	↑	↓↓	↓
<i>Practice + Explicit Feedback Condition</i>	↓	↑	↓	↓

TABLE 5.2: Visual summary of model findings reported in Table 5.1. Upwards arrow (rose-colored cell) indicates increases in the measure value after the training. Downwards arrow (blue cell) indicates decreases in the measure value after the training, and two downwards arrows (darker blue cell) indicates greater decreases in the measure value in the repeated practice condition compared to when explicit feedback was shown alongside repeated practice.

(e.g., in [62, 95]). As shown in Figure 5.9, for cases in which the past experienced worker was inaccurate with respect to the model's predictive target (left), around 60% of participants at both the start and end of the training made decisions that were accurate with respect to the past experienced worker. However, for cases in which the past experienced worker was accurate with respect to the model's predictive target, only around 40% of participants made decisions that were accurate with respect to the past experienced worker in the first case. Yet, by the 24th case, this number increased to around 60% of cases, erasing any differences in decision behavior for cases in which past experienced workers were accurate versus inaccurate with respect to the model's predictive target. As shown in Table 5.1, explicit feedback did not have a significant effect on Worker-based Accuracy.

5.4.3 Guessing Error: Participants learn to make improved predictions of the AI model's behavior

- (Finding-5) **Participants improved in their ability to predict the AI model's behavior after repeated practice making AI-assisted decisions.**
- (Finding-6) **Explicit feedback led to lesser improvement in participants' ability to predict the AI model's behavior, compared to practice alone.**

While disagreeing with the AI prediction more often with time, participants were simultaneously *getting better at guessing what the AI prediction would be on a given case*. Participants who received practice alone, compared to those who additionally received feedback on their decisions, saw somewhat more improvement in their guessing performance. As shown in Table 5.1, there was a significant decrease in guess mean squared error in the post-assessment as compared to the pre-assessment (Coef.=−3.03; $p < .001$; 95% CI: [−3.45, −2.60]). Additionally, participants assigned to the *Practice* condition saw a greater improvement in their guessing performance as compared to participants in the *Practice + Explicit Feedback* condition (Coef.=0.608; $p < .05$; 95% CI: [0.05, 1.15]). Overall, this result indicates that, despite disagreeing with predictions, participants in both conditions were actively engaging with the learning activity and improving their skills at mentally simulating the behavior

of the model. However, participants who did *not* receive explicit feedback on their AI-assisted decisions improved these skills more. As we discuss in Section 5.5.2, it is possible that, compared to the explicit feedback, the information reported in the case (including the qualitative case narratives) served as a more powerful signal through which to learn how the AI model behaves on different cases, compared to the explicit feedback signals provided. It is possible that the *Practice* condition saw greater improvement because explicit feedback on AI decisions distracted participants from learning to predict model behavior based on qualitative narratives.

5.4.4 Process Analyses: Participants learn to make decisions more like experienced workers

To understand participants' decision-making process, we qualitatively examined their open text responses. These included (1) participants responses in each of the pre- and post-assessments, where they explained how they made their AI-assisted decisions and (2) participants' explanations at the end of the study, describing what goals guided their decisions.

- (Finding-7) **Participants' explanations for their decisions reference their use of model unobservables to help them make AI-assisted decisions.**
- (Finding-8) **Participants' explanations of their decision-making goals indicate they target ensuring immediate safety to the child when making AI-assisted decisions.**

Informing decisions using qualitative case narratives.

An examination of participants' explanations of how they made their AI-assisted decisions in the pre-assessment and post-assessment phases of the study provide evidence of how they relied on the qualitative case narratives to inform their decisions. Much like how actual experienced social workers used AI-based decision support tools in their day-to-day jobs, participants drew on context-specific details from the allegations to inform their decisions. For example, participants sometimes drew on their interpretations of the qualitative case narratives, to inform decisions to disagree with the AI prediction. When explaining their decision to screen out a case with an AI prediction of 16, one participant described: "Despite the high risk score[,] I do not see evidence of maltreatment based on the details. The parent was angry at a third party and may have roughly grabbed the child but that does not rise to maltreatment unless it can be proven that there is a pattern of rough handling."

Similarly, another participant considered missing information in the allegation (e.g., about intentions and causes) to appropriately weigh the severity of the reported allegation, in comparison to the AI prediction: "Although the risk score is 12, the child appears to be physically healthy and has a positive mood. There are many reasons a knife could be in a bedroom, and it is possible the kid's struggles socially and academically could be caused by a learning disability." In another case, a participant explained a decision to screen in a case, despite a low AI prediction, given allegations reporting violent behavior: "Risk score appears too low for this case. Multiple examples of violence brought on by behavioral problems of the mother in the home, as well as child acting out at school, all meet definition of legal maltreatment."

In other cases, participants referenced the allegations to gain more insight into the child's current living situation, and whether there are signals indicating there

may be an upwards or downwards trend for improvement. For example, in explaining their decision to screen out a case, the participant described how a family appears to be on a path towards recovering from past challenges: “The mother has a support system in place. She has the grandmother, and is receiving public benefits/services. Her treatment team reports that she should be able to fully recover and care for her child. And the child is currently clean and well cared for.” In another case, a participant noted the lack of support and resources that the child currently has, alongside the history of substantiated referrals, to inform their decision to screen in a case: “The fact that the child is not taking her prescribed medication, not receiving counseling for her mental condition, and at risk of homelessness demonstrates that the child is in severe risk of physical and emotional harm. Moreover, the substantiation of past referrals lends further credibility and urgency to the current one.”

Overall, participants did not mention the AI prediction as often when explaining their decisions. In particular, participants mentioned using a “score” (i.e., the “AI Risk Score” shown in the case referral) in 8.72% of the explanations (for 247 out of 2832 total cases). Explanations considering information from a “score” decreased after the Training Activity Phase, corroborating findings that participants relied on the AI prediction less with increased practice. In the pre-assessment Phase (cases 1 - 4), participants considered the “score” for 194 cases (79% of the 247 mentions). In the post-assessment Phase (cases 21 - 24), the frequency dropped to only 57 cases (21% of 247 mentions). In many of these cases, participants referenced the score as one of multiple information sources they drew on to explain their decision (e.g., “The Risk score is a 20, there is commotion and drug use within the family.”). However, in 17 of the 57 cases that mentioned a “score” in the post-assessment Phase, participants were describing why they were disagreeing with it (e.g., “even if the risk score is fairly low, the situation doesn’t seem to be very under control and could break any moment”). The 247 mentions of a “score” across the pre- and post-assessments were from 112 total participants, split roughly evenly across the *Practice* and *Practice + Explicit Feedback* conditions.

Making decisions with the goal of ensuring immediate safety.

Beyond leveraging their unique ability to access and integrate context-specific information when making decisions, the goals that participants had when making decisions may also explain the learning effects. In particular, it is possible that participants learned to disagree with the AI prediction, while learning to make decisions like past experienced workers, partly because their own decision-making goals aligned more strongly with those of the past experienced workers. At the end of the experiment, participants were asked to describe their own goals when making child maltreatment screening decisions. Many participants described the importance of considering near-term safety risks or harms to the child, which is misaligned with the longer-term outcomes that the AI model is trained to predict. Both participants with and without domain knowledge in social work described goals related to “immediate safety and harm,” aligning with the goals documented of the actual past experienced social workers in this domain. For example, one social worker described: “I am looking for immediate safety concerns first. Then making sure physical needs are being met. If either of these are a concern, it is an immediate screen in.” Another social work graduate student described the importance of “current” evidence: “I assess first for evidence of maltreatment or harm currently going on in the home and then consider if additional investigation is necessary to rule out possible abuse or

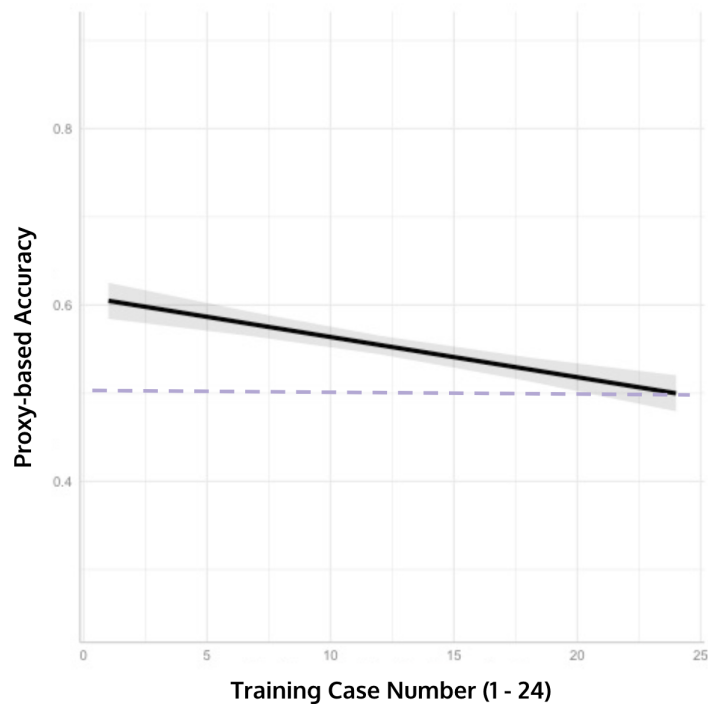


FIGURE 5.10: Participants' Proxy-based Accuracy measures against the baseline accuracy of the model (in purple).

neglect, I then consider if there is enough "probable cause" to truly investigate this further." A participant without social work domain expertise similarly described: "My goals were whether or not there would be any immediate danger posed to the child either through their environment or by their parents."

[t]

5.4.5 Proxy-based Accuracy: Participants become *less* aligned on the model's targeted proxy

- (Finding-9) **Participants' decisions become less accurate with respect to the proxy outcome targeted by the AI model after repeated practice.**
- (Finding-10) **No effect of explicit feedback on the accuracy of participants' decisions with respect to the proxy targeted by the AI model.**

Participants gradually began to disagree with the AI prediction regardless of whether the AI was inaccurate or accurate with respect to the model's predictive target. As shown in Figure 5.8, for the first case in which the AI was accurate with respect to the model's predictive target, 80% of participants agreed with the AI prediction. For the first case in which the AI was inaccurate with respect to the model's predictive target, 60% of participants agreed with the AI prediction. However, in the 24th case, only 40% of participants agreed with the AI prediction, regardless of whether the AI prediction was accurate or inaccurate with respect to the model's predictive target.

Importantly, this gradual disagreement with the AI prediction occurred *regardless of whether the participant received feedback on their decisions*. As shown in Table 5.1, there is no feedback effect on participants' decisions to agree or disagree with the AI prediction over time (Coef. 0.02, $p < .50$, 95% CI: [-0.03, 0.09]).

5.5 Discussion

In this paper, we define a notion of appropriate reliance called *critical use*, which emphasizes human decision-makers' ability to situate AI predictions against knowledge that is uniquely available to them but unavailable to the AI model. Through an experimental investigation of how training can support critical use of AI predictions in making AI-assisted child maltreatment screening decisions, we explore the effects of repeated feedback and explicit feedback on *what* and *how* participants learn. In the following, we discuss interpretations of our results and avenues for future work.

5.5.1 Evidence of *critical use* from training

Our findings suggest that training improved *critical use* of the AI tool—humans' ability to situate AI predictions against potentially complementary knowledge available to them (but not the AI model). A body of prior research on AI-assisted child maltreatment screening has documented how social workers draw on their knowledge of qualitative, contextual information to calibrate their reliance upon AI-based decision support (e.g., [221, 359, 363]). In doing so, experienced workers can overcome some of the challenges posed by target-construct mismatch and limitations in the information available to AI models, mitigating erroneous AI outcomes and reducing disparities in decisions [62, 95]. In our study, we found that **participants learned to calibrate their reliance upon AI predictions in ways that resembled experienced workers**. In particular, with increased practice, participants learned to disagree more often with AI predictions (Section 5.4.1) and, in turn, came to make decisions that aligned more with those of past workers with extensive experience making AI-assisted decisions (Section 5.4.2). Moreover, through practice making AI-assisted decisions participants improved their ability to mentally simulate the AI model and predict how it would behave on specific cases (Section 5.4.3). Our qualitative examination of participants' explanations suggest that participants drew upon knowledge regarding the actual decision-making goals (i.e., screening for shorter-term safety) and interpreted qualitative narratives unavailable to the AI model in order to inform their AI-assisted decisions (Section 5.4.4) and learn how the AI model behaves on different kinds of cases. However, we note that future work is needed to understand the *extent* to which specific downstream impacts of target-construct mismatch and information asymmetries (e.g., systematic errors and unfairness in decision-making) can be mitigated through training.

Analyses using a standard measure of accuracy for human-AI decision-making would indicate that increased practice led participants to make *less* accurate decisions, when accuracy is measured with respect to the model's target proxy (Section 5.4.5). Interpreting proxy-based accuracy in isolation from the rest of our analyses may lead to the conclusion that participants are simply falling into patterns of algorithm aversion [53, 102]. However, in analyzing indicators of *critical use*, including both process- and outcome-oriented signals, our findings suggest that participants disagreed with AI outputs in sophisticated ways. Indeed, in complex, real-world decision-making settings, common outcome-focused metrics for evaluating AI-assisted decisions—for example, comparing human decisions with the AI model's ground truth label—can provide an incomplete understanding of the quality of each decision [149, 221]. Our findings point to a compelling opportunity to improve evaluations of AI-assisted decision-making in other complex domains, by designing and employing context-specific measures of *critical use*, *in addition to examining proxy-based accuracy*. Moreover, to expand the set of indicators that could

be used to measure decision-making, future work should innovate on ways to better capture process-based signals for decision quality (e.g., [249, 285]). We further discuss how our study findings and approach may generalize across domains in Subsection 5.5.4.

Limitations For the purposes of our study, we selected cases where the AI tool had an average proxy-based accuracy of (50%). This is lower than the the AFST model’s actual accuracy with respect to its proxy. Our decision to include cases that aggregated to 50% proxy-based accuracy of the AI tool was informed by our goal of having a well-balanced set of examples to conduct case type-specific analyses and of prior literature discussing the risks of over-reliance in the absence of sufficient examples of the AI model erring [320]. While we cannot rule out that this contributed to increased disagreement with the AI tool, our findings indicate that it is *not sufficient* to explain participants’ learning. For example, participants aligned less with the model proxy following training even in the *Practice* condition, when participants did not receive explicit feedback on the accuracy of the AFST with respect to its targeted proxy (see [271]).

5.5.2 Why did explicit feedback on AI-assisted decisions not enhance learning?

We did not observe significant impacts of showing explicit feedback on participants’ *decision-making*, compared with practice alone. Furthermore, although the training improved participants’ ability to predict how the AI model would behave on a given case, participants who received practice opportunities *without* explicit feedback saw greater improvements than those who received feedback. Taken together, our findings suggest that when it comes to learning to make AI-assisted decisions, the qualitative case narratives participants see and interpret (which are inaccessible to the AI model) may have had a bigger impact than we had originally anticipated. These narratives, which were present in both conditions, served as a rich information source regarding the plausibility of individual AI predictions, through which participants were able to learn to calibrate their reliance on AI predictions. Thus, even in the absence of *explicit* feedback on decisions, a training interface that provides accelerated practice, with opportunities to cross-check AI predictions against qualitative narratives—a form of *implicit* feedback on the reliability of individual AI predictions—appears to be a stronger baseline condition than expected. Our findings further suggest that participants who were shown explicit feedback may not have perceived the feedback provided as useful signals of decision quality, relative to the signals provided through the qualitative case narratives.

Prior work from the learning sciences suggests that for certain learning tasks, providing learners with “grounded” feedback—concrete representations that offer rich, meaningful signals about whether or not an individuals’ targeted outcome is achieved—is more effective than simply telling learners whether a response is correct or incorrect. For example, when teaching students to perform algebraic transformations, showing students a graphical representation of equations they enter can help students immediately *recognize for themselves* whether an equation is likely to be correct or incorrect[450]. Future work on training for AI-assisted decision-making should further explore the design space of rich, grounded feedback mechanisms that can help humans (learn to) calibrate their reliance. In particular, whereas our study provided explicit feedback in the form of categorical outcomes and workers’ final decisions, future work should explore what forms of grounded feedback

could support more effective process-oriented learning in a given context. In complex decision-making settings such as child maltreatment screening, observable outcomes and even final decisions) are noisy signals for decision making quality [62, 222]. Therefore, rather than providing human decision-makers with outcome-based feedback that attempt to directly “tell” a learner whether a given decision was accurate, providing rich, grounded feedback may make it easier to engage learners’ sensemaking and help them *assess for themselves* whether and why a decision may be right or wrong. Training that focuses on better scaffolding the decision-making process—for example, by pairing a novice decision-maker with an expert decision-maker to collaboratively reason about each decision—may improve critical thinking and sensemaking processes that also allow humans to better calibrate reliance on ADS outputs.

5.5.3 On the absence of domain expertise effects

We found that regardless of participants’ level of self-reported domain expertise in social work, they learned to disagree with AI predictions in a manner that resembled the disagreement patterns of experienced workers. Participants with greater self-reported domain expertise did not disagree more with the AI prediction or make more accurate decisions with regards to the past experienced worker, compared with other participants.

We do not view the absence of domain expertise effects in our study as evidence that domain expertise has zero impact on how people learn to make AI-assisted decisions. In our study, “domain expertise” was broadly defined to include individuals with domain knowledge in *any* area of social work, not limited to knowledge of child welfare or experience with child maltreatment screening. Participants’ roles included current or former social workers and social work graduate students. We had originally hypothesized that broad knowledge in social work would influence how and what participants learn through a training for AI-assisted child maltreatment screening. However, given the specialized nature of this task, it is very plausible that a tighter recruitment criterion—for example, requiring “domain expert” participants to have prior domain knowledge in child welfare or even prior experience with child maltreatment screening—would reveal more substantial differences between participants with more or less domain expertise. In addition, it is possible that the specific metrics used in our analyses were unable to capture differences in learning and decision-making between the “domain experts” in our study and other participants. Finally, a third possibility is that general human abilities shared by non-domain-experts (i.e., having access to and being able to interpret qualitative information, and having knowledge of broader decision goals) are sufficient on their own to support learners in approximating experienced workers patterns of disagreement with AI predictions.

We emphasize that our results do not indicate that participants learned to make decisions *as well as* experienced workers during the course of our training activity. Rather, our findings indicate that participants learned to predict how the AI model would behave on different cases and learned to disagree with the AI in a manner resembling experienced workers. Future work is needed to tease apart the three potential explanations above.

5.5.4 Generalizability of design decisions and findings

As a concept, *critical use* is applicable to any domain where humans and AI models have access to complementary knowledge. Critical use emphasizes that humans can be better supported in calibrating reliance on AI-based decision-making tools, by learning to leverage complementary knowledge they have as human decision-makers (e.g., knowledge of additional decision-relevant features, or knowledge of the true objectives of a decision-making task). We expect that several of the findings and implications discussed above—for example, the importance of providing learners with concrete, domain-grounded forms of feedback against which they can cross-check AI outputs—will generalize to other domains, beyond the context studied in the current research. However, we note that the indicators of critical use that we adopted in our study are highly specific to the child maltreatment screening context. As discussed above, our choices of indicators were informed by prior empirical research studying how experienced workers in this domain calibrate their reliance upon AI predictions; this specific set of indicators and the interpretations adopted in this study may not generalize well to other domains. Accordingly, our goal in this research is not to make any universal recommendations on which specific, measurable decision-making outcomes and learning indicators are “good” or “bad” across domains. Rather, we emphasize that the approach of drawing upon prior knowledge to design domain-specific training interfaces and indicators of critical use can be generalized to inform the design of improved training materials and learning measures in other complex, real-world domains. Future work is need to explore how our approach can be adapted to other domains, and to investigate what training designs and measures of critical use are most appropriate in different decision-making settings.

Chapter 6

Improving AI Adoption and Governance

This chapter is taken from the following paper:

Anna Kawakami, Amanda Coston, Haiyi Zhu, Hoda Heidari, and Kenneth Holstein. “The Situate AI Guidebook: Co-Designing a Toolkit to Support Multi-Stakeholder, Early-stage Deliberations Around Public Sector AI Proposals.” In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.

Abstract. Public sector agencies are rapidly deploying AI systems to augment or automate critical decisions in real-world contexts like child welfare, criminal justice, and public health. A growing body of work documents how these AI systems often fail to improve services in practice. These failures can often be traced to decisions made during the early stages of AI ideation and design, such as problem formulation. However, today, we lack systematic processes to support effective, early-stage decision-making about *whether* and *under what conditions* to move forward with a proposed AI project. To understand how to scaffold such processes in real-world settings, we worked with public sector agency leaders, AI developers, frontline workers, and community advocates across four public sector agencies and three community advocacy groups in the United States. Through an iterative co-design process, we created the *Situate AI Guidebook*: a structured process centered around a set of deliberation questions to scaffold conversations around (1) *goals and intended use* for a proposed AI system, (2) *societal and legal considerations*, (3) *data and modeling constraints*, and (4) *organizational governance factors*. We discuss how the guidebook’s design is informed by participants’ challenges, needs, and desires for improved deliberation processes. We further elaborate on implications for designing responsible AI toolkits in collaboration with public sector agency stakeholders and opportunities for future work to expand upon the guidebook. This design approach can be more broadly adopted to support the co-creation of responsible AI toolkits that scaffold key decision-making processes surrounding the use of AI in the public sector and beyond.

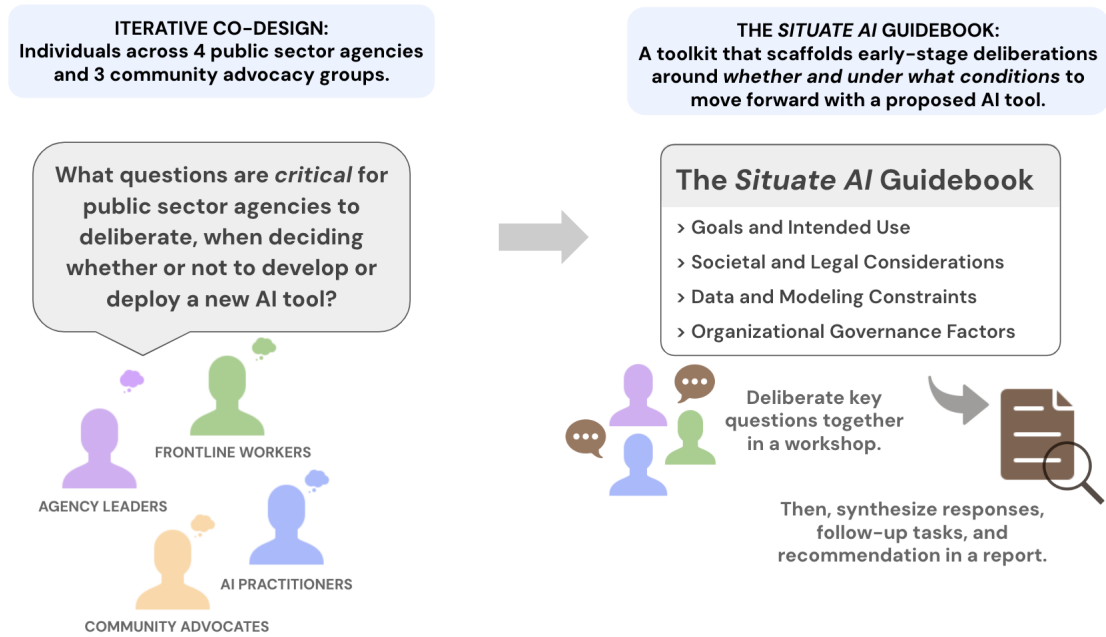


FIGURE 6.1: We conducted co-design activities and semi-structured interviews with public sector agency workers (agency leaders, AI practitioners, frontline workers) and community advocates to understand the questions they believed were *critical to discuss* yet currently overlooked before deciding to move forward with a public sector AI proposal. The *Situate AI Guidebook* synthesizes these key considerations into a toolkit to scaffold early-stage deliberations around *whether* and *under what conditions* to move forward with developing or deploying a proposed public sector AI tool.

6.1 Introduction

Public sector agencies in the United States are rapidly adopting AI systems to assist or automate services in settings such as child welfare, credit lending, housing allocation, and public health. These tools have been introduced to help overcome resource constraints and limitations in human decision-making [74, 261]. However, as a growing body of work has documented, public sector AI tools have often failed to produce value in practice, instead exacerbating existing problems or introducing new ones [62, 221, 303, 443]. For example, the Michigan Unemployment Insurance Agency developed an AI-based fraud detection system (MiDAS); the agency stopped using the tool after realizing it falsely flagged over 90% of its cases—a discovery that was made only *after* the tool had been in use for over two years, impacting hundreds of thousands of people along the way [21]. Similarly, following the deployment of an AI-based tool for child maltreatment screening, Allegheny County’s Department of Human Services faced significant criticism after the tool was found to exacerbate biases against Black and disabled communities [119, 138, 163, 164]. Research indicates that these problems in deployment were a consequence of fundamental conflicts between the tool’s design on the one hand, and data limitations and worker needs on the other [119, 138, 221, 222]. Many other public sector agencies have dropped deployed AI tools for similar reasons, even after investing significant resources into their development (e.g., [206, 355]).

Many failures in public sector AI projects can be traced back to decisions made during the earliest problem formulation and ideation stages of AI design [80, 319,

443, 463]. AI design concepts that make it to production may be “doomed to fail” from the very beginning, for a variety of reasons. For example, AI design concepts have often been conceived in isolation from workers’ actual decision-making tasks and challenges, leading to AI deployments that are not actually viable in practice [178, 222, 427, 459, 463]. Similarly, teams often propose design concepts for new tools that cannot possibly be implemented in an effective, safe, or valid way given technical constraints, such as the availability and quality of data [80, 336, 443, 463]. However, discussion of such constraints is commonly left to later stages of the AI lifecycle, by which point teams have invested in an idea and may be more reluctant to explore alternative ideas [221, 463]. While agencies utilizing AI may be motivated to try to mitigate issues at later project stages, such attempts are unlikely to yield meaningful improvements if fundamental issues around the problem formulation and solution design are left unaddressed [138, 222, 395, 443, 463].

In this paper, we ask: **How can we support public sector agencies in deciding whether or not a proposed AI tool should be developed and deployed in the first place?** Today, we lack systematic processes to help agencies make informed choices about which AI project ideas to pursue, and which are best avoided. As AI tools proliferate in the public sector, the failures discussed above indicate that agencies are repeatedly missing the mark with AI innovation. While existing responsible AI toolkits have provided guidance on ways to support AI development and implementation to ensure compliance with the relevant principles and values (e.g., [132, 279, 291, 342]), most existing toolkits are designed for use in industry contexts. Furthermore, most toolkits start from the assumption that the decision to develop a particular AI tool has *already* been made.

To address these gaps, we introduce the *Situate AI* Guidebook: a toolkit to scaffold early-stage deliberations around *whether and under what conditions* to move forward with the development or deployment of a proposed public sector AI innovation. To ensure that our guidebook and process design is informed by existing organizational needs, practices, and constraints in the public sector, we partnered with 32 individuals, spanning a wide range of roles, across four public sector agencies and three community advocacy groups across the United States. Over the course of 8 months, we iteratively designed and validated the guidebook with a range of stakeholders, including (1) public sector agency leadership, (2) AI developers, (3) front-line workers, and (4) community advocates. The public sector agencies we partnered with represent different levels of experience and maturity with AI development and deployment: At the time of this research, some had just begun ideating ways to integrate AI tools into their agencies’ processes; some were already in the process of developing new AI tools; and some had already experienced failures in AI tool deployment that led to halts in their use. The community advocacy groups include organizations that, among other areas of focus, represent and support community members in navigating challenging interactions with public services (e.g., parents negatively impacted by the child welfare system).

We conducted formative semi-structured interviews and iterative co-design activities that guided the content and process design of the *Situate AI* Guidebook. In particular:

- Through semi-structured interviews, we developed an understanding of public sector agencies’ current practices and challenges around the design, development, and evaluation of new AI tools, in order to identify opportunities for new processes to improve current practice.

- Through co-design activities, participants ideated and iterated upon a set of questions that they believed were critical to consider before deciding to move forward with the development of a proposed AI tool. In addition, they described how they envisioned a deliberation process could be effectively structured for adoption at their agencies.

The resulting set of deliberation questions spanned a broad range of topics, from centering community needs to surfacing potential agency biases, given their positionality—topics which are relatively understated in existing Responsible AI toolkits developed for industry contexts. Notably, participants gravitated toward deliberation questions that promoted reflection on potential differences in perspective among the various stakeholders of public sector agencies (e.g., agency workers, frontline workers, impacted community members), surrounding topics such as the problem to be solved by an AI tool, notions of “community”, or understandings of what it means for decision-making to be “fair” in a given context. This work presents the following contributions:

1. **The *Situate AI Guidebook*¹ (Ver.1.0)** – the first toolkit co-designed with public sector agencies and community advocacy groups to scaffold *early-stage deliberations* regarding *whether or not* to move forward with the development of a proposed AI tool.
2. **A set of 132 co-designed deliberation questions** spanning four high level topics, (1) *goals and intended use*, (2) *societal and legal considerations*, (3) *data and modeling constraints*, and (4) *organizational governance factors*. Participants indicated these considerations are *critical to discuss* when deciding to move forward with the development of a proposed AI tool, yet are not proactively or deliberately discussed today.
3. **Guidance on the overall decision-making process** that the *Situate AI Guidebook* can be used to support, informed by how participants envisioned they would use the guidebook in their agencies and by prior literature discussing related challenges that threaten the practical utility of research-created artifacts [279, 453].
4. **Success criteria for using the guidebook** informed by participants’ existing challenges, prior literature, and signals that participants themselves described as valuable in assessments regarding the guidebook’s ability to promote meaningful improvements in their agency.

In the following sections, we first overview relevant bodies of prior literature to help ground and motivate the creation of our toolkit (Section 7.11). We then describe the approach we took to collaboratively develop the *Situate AI Guidebook* (Section 7.12), and describe the guidebook’s major components, including its guiding design principles (Section 6.4.1), deliberation questions (Section 6.4.2), process design (Section 6.4.3), and success criteria (Section 6.4.4). We conclude with a discussion of anticipated challenges, as well as directions for future research aimed at understanding how to implement such deliberation processes most effectively. We also discuss implications for future co-design of responsible AI toolkits intended to promote meaningful change in public sector contexts (Section 7.14). The public sector agencies we partnered with in this study plan to explore the use of the guidebook through pilots, to identify further avenues for improvement.

¹<https://annakawakami.github.io/situateAI-guidebook/>

6.2 Background

6.2.1 Public Sector AI and Overcoming AI Failures

In the United States, public sector agencies are government-owned or affiliated organizations occupying the federal, state, county, or city government, responsible for making decisions around the allocation of educational, welfare, health, and other services to the community [299]. Public sector agencies across the United States are exploring how to reap the benefits of AI innovations for their own workplaces. AI tools promise new opportunities to improve the efficiency of public sector services, for example, by increasing decision quality and reducing agency costs [30, 73, 317, 434]. In 2018, 83% of agency leaders indicated they were willing or able to adopt new AI tools into their agency [3]. In the public sector, there is also a recognition that developing AI tools in-house can help ensure that they are better tailored to meet agency-specific needs, ensure they are trained on representative datasets, and account for local compliance requirements [115]. However, achieving responsible AI design in the public sector has proven to be an immense challenge [143, 145, 354, 439]. The domains where agencies are attempting to apply AI are often highly socially complex and high-stakes—including tasks like screening child maltreatment reports [361], allocating housing to unhoused people [247], predicting criminal activity [235], or prioritizing medical care for patients [310]. In these domains, where some public sector agencies have a fraught history of interactions with marginalized communities [2, 344], it has proven to be particularly challenging to design AI systems that avoid further perpetuating social biases [62], obfuscating how decisions are made [222], or relying on inappropriate quantitative notions of what it means to make accurate decisions [80]. Public sector agencies are increasingly under fire for implementing AI tools that fail to bring value to the communities they serve, contributing to a common trend: AI tools are implemented then discarded after failing in practice [142, 354, 459, 463].

Research communities across disciplines (e.g., HCI, machine learning, social sciences, STS) are beginning to converge toward the same conclusion: **Challenges observed downstream can be traced back to decisions made during early problem formulation stages of AI design.** Today, we lack concrete guidance to support these early stages of AI design [80, 319, 443, 463]. For example, after observing decades of failures to develop clinical decision support tools that bring value to clinicians, researchers have found that AI developers may **lack an adequate understanding of which tasks clinicians desire support for**, leading to the creation of tools that target problems that clinicians do not actually have [113, 142, 459]. These trends are beginning to surface across other domains that have more recently begun to explore the use of new AI tools. In social work, researchers have found that technical design decisions in decision support tools reflect **misunderstandings around the type of work that social workers actually do**, leading to deployments where, for instance, the underlying logic of the model conflicts with how workers are trained and required to make decisions [221, 222]. Others have surfaced how seemingly technical design decisions made during early stages of model design actually embed **policy decisions that conflict with community values and needs** [138, 362, 395].

In addition to concerns regarding how well the problem formulation and design of a given AI tool reflects worker practices and community values, there is a concern that AI tools deployed in complex, real-world domains may be **conceived without adequate consideration for the actual capabilities of AI**. For example, examining a range of real-world decision support tools (e.g., in criminal justice, child welfare, tax

lending), researchers have argued that existing AI deployments lack validity, due to limitations in the types of data that can be feasibly collected to train the desired model [80, 303, 336, 443]. This highlights the need for developers and organizations to reflect upon technical constraints and limitations at earlier stages of the AI development lifecycle, such as when evaluating whether or not to pursue a proposed AI project in the first place.

While public sector agencies have emphasized the potential to improve decision accuracy and reduce bias as a key motivation to use new AI tools (e.g., [89]), these challenges around the problem formulation and design of AI systems implicate the veracity of these claims [80, 222, 336, 395, 443, 463]. We identify a significant opportunity to better support public sector agencies in making systematic, deliberate decisions regarding whether or not to implement a given AI tool proposal. Given the vast potential for harm, and the similarly vast potential for AI systems to meaningfully support workers and improve services in the public sector, it is critical to support agencies through concrete guidance and processes in making more informed decisions around which AI tool proposals to pursue, and which to avoid.

6.2.2 Toolkits for Responsible AI Governance

In an effort to support responsible design and development of AI systems in practice, the HCI, ML and FAccT research communities have contributed a range of responsible AI toolkits. These toolkits are intended to support and document assessments of AI systems, including their (potential) impacts (e.g., [341]), intended use cases (e.g., [291]), capabilities and limitations (e.g., [342]), dataset quality (e.g., [136, 332]), and performance measures (e.g., [291]). Many of these toolkits are intended to be used as communication tools. For example, Model Cards provide a structure for communicating information regarding the intended uses, potential pitfalls, and evaluation measures of a given ML model, to support assessments of suitability for a given application and context of use [291]. Recent research surveying these toolkits have found that the majority of existing toolkits frame the work of AI ethics as “technical work for individual technical practitioners [453]. For the majority of existing responsible AI toolkits, the primary users are ML practitioners, limiting the forms of knowledge and perspectives that inform the work of “AI ethics” [453]. A smaller number of toolkits have been designed for use by organization-external stakeholders, to support impacted end-users in interrogating and analyzing deployed automated decision systems (e.g., the Algorithmic Equity Toolkit [240]); provide impacted stakeholders with an opportunity to share feedback on an AI system’s use cases and product design (e.g., Community Jury [76]); or support philanthropic organizations in vetting public sector AI technology proposals [1]). In all of these examples, the toolkit supports examinations of AI systems that have already been developed and sometimes even deployed.

Most existing toolkits assume that the decision to develop a particular AI system has already been made. Therefore, even when they are intended to support reflection and improvement of the AI system, the types of improvements that could stem from using the toolkit tend to be limited to those that would not require fundamental changes to the underlying technology. Meanwhile, while some existing responsible AI toolkits target earlier stages of AI development (e.g., [464]), these have primarily been designed for *private sector* contexts. Yet there is good reason to expect that public sector agencies would benefit from tailored responsible AI tools. For instance, compared with the private sector, there is a greater expectation that public sector agencies exist to serve people and are expected to make decisions that

center communities' needs. When making decisions as critical as what new AI tools to deploy, agencies are expected to adhere strongly to values such as deliberative decision-making, public accountability, and transparency. To date, there exists minimal concrete and actionable guidance on how to support *public sector agencies* in scaffolding *early-stage* deliberation and decision-making.

A related existing artifact is the AI Impact Assessment, described as a “process for simultaneously documenting an [AI] undertaking, evaluating the impacts it might cause, and assigning responsibility for those impacts” [294]. AI Impact Assessments have been proposed for both public and private sector contexts, and are intended to be completed either at an early stage of AI design (e.g., [9]), or after an AI system is developed or deployed (e.g., [284]). Another related artifact is the Data Ethics Decision Aid (DEDA) [132], a framework to scaffold ethical considerations around data projects proposed in the Dutch Government. However, neither of these examples are designed as *deliberation toolkits*, to promote collaborative reflection and discussion around the underlying problem formulation or solution design of an AI tool. AI Impact Assessments and ethical decision aids have also not typically been designed in collaboration with the stakeholders they intend to serve. With recent research suggesting low adoption of responsible AI toolkits in real-world organizational contexts [337, 453], a co-design approach with organizational stakeholders has the potential to generate responsible AI tools that work in practice.

6.3 Methods: Co-Design and Validation of the *Situate AI Guidebook*

To iteratively co-design and validate the *Situate AI Guidebook*, we conducted semi-structured interviews and co-design activities with a range of stakeholders both within and outside of public sector agencies. In this section, we describe participants' backgrounds, the approach and resources used in our iterative co-design process, and our data analysis approach.

6.3.1 Participants and Recruitment

We co-designed the *Situate AI Guidebook* with individuals from **four public sector agencies** across the United States. Collectively, this set of public sector agencies has experienced a range of decision-making scenarios around the creation or use of AI-based decision tools. All four agencies are currently ideating new forms of AI-based tools, three have already implemented AI tools, and at least one had previously deployed an AI tool and subsequently decided to abandon it. From these agencies, we wanted to include stakeholders at different levels of the organizational hierarchy including those with experience making relevant decisions and those who are involved in the development or consumption of AI tools but who are not typically involved in decisions around development and deployment. We therefore included participants from three core stakeholder groups: 1) **Agency leaders (L)** who are in director or managerial roles, typically involved in agency- or department-level decisions including whether to design and deploy a particular AI tool, 2) **AI developers, analysts, and researchers (A)** who are in development, analysis, or research teams internal to a given public sector agency and typically build and evaluate AI tools, and 3) **Frontline decision-makers (F)** whose occupations bring them in direct contact with the community their agency serves and whom an AI tool may be intended to assist. Because we wanted to learn from additional frontline decision-makers but

had access only to a limited number at the public sector agencies we connected with, we recruited additional participants beyond these agencies, with relevant professional backgrounds. These included social work graduate students with prior field experience making frontline decisions in public sector agencies.

In addition, we co-designed the guidebook with individuals from **three community advocacy groups** across the United States, including family representation and child welfare advocacy groups. Individuals from these organizations created the fourth stakeholder group: 4) **Community advocates (C)** who represent and meet community members' needs around public services. While the *Situate AI Guidebook* is intended to be used by workers within a public sector agency, we included community advocates because we wanted the guidebook to represent their perspectives regarding the most critical considerations for moving forward with an AI tool design. As discussed in Section 6.4.3, we also worked with community advocates to begin envisioning what a future version of the toolkit, aimed at engaging community members in the deliberation process, might look like.

In total, 7 agency leaders; 7 developers, analysts, and researchers; 7 frontline decision-makers; and 11 community advocates participated in the co-design process. To recruit public sector agencies, we contacted 19 U.S. public sector agencies at the state, city, or county level with human services departments. We received responses from five agencies. Following a series of informal conversations to share our research goals and study plans, four of the agencies decided to participate in the study. To recruit individuals from community advocacy organizations, we contacted community leaders and advocates across 8 organizations. While we requested individual study participation, some participants preferred to participate in the research study in small groups. By participating in groups, they believed they could provide a more extensive set of insights together. 21 out of 25 sessions were conducted individually, and the remaining four were group interviews. For ease of communication, we will use the singular noun "participant" throughout the remainder of the paper.

6.3.2 Iterative Co-Design and Validation

The *Situate AI Guidebook* integrates findings across semi-structured interviews and co-design activities, which were conducted over the course of eight months between November 2022 and June 2023. The study sessions were ~90 minutes long for public sector workers, who were involved in both the interviews and co-design activities; the study sessions were ~60 minutes for community advocates, who were only involved in the co-design activities.

Formative Semi-structured Interviews

To ensure that the *Situate AI Guidebook* is designed to address real-world needs and goals, we conducted semi-structured interviews with public sector agency workers to understand (1) their existing challenges and barriers to making decisions around AI systems and (2) desires for improving their current decision processes. Specifically, to understand existing decision-making processes, we asked each participant to recall a specific prior experience in which they or their agency needed to decide whether to move forward with the development or use of a new AI-based tool. As participants shared their stories, we asked follow-up questions to probe on possible causes behind the challenges that they described. For example, after describing how they previously made a related decision, we asked "What's challenging

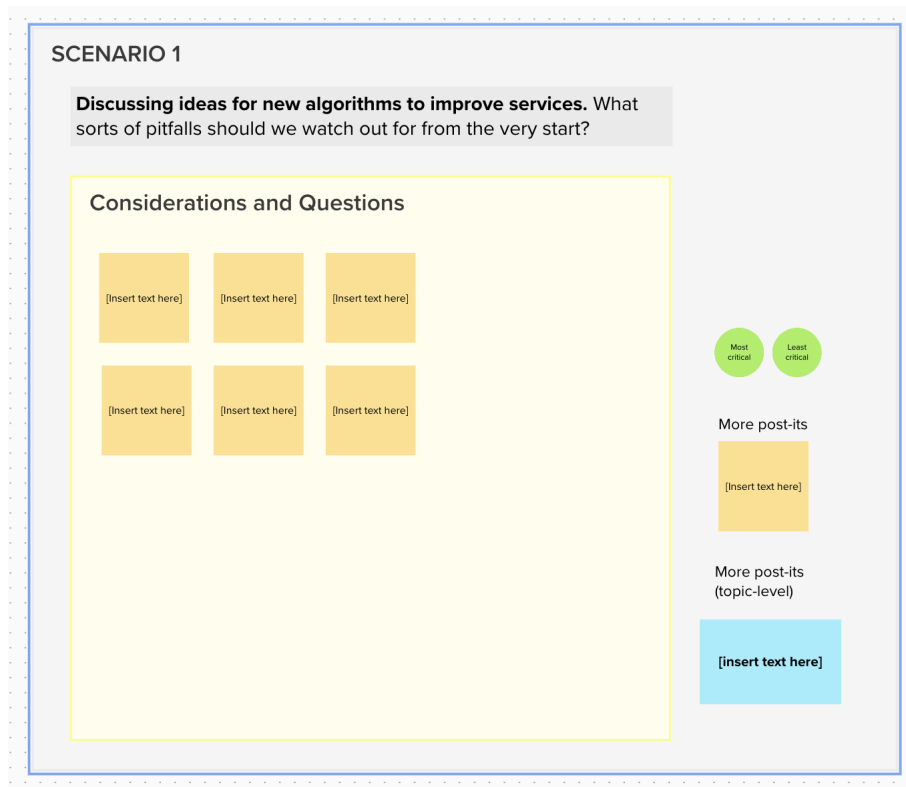


FIGURE 6.2: Screenshot of a blank board shown to participants at the start of the co-design activity on Mural. This board presents Scenario 1: Discussing ideas for new algorithms to improve services.

to do well now, when you’re making those decisions?” or “What would you ideally want to discuss in conversations surrounding those decisions?” If a participant shared that they had not personally been involved in decisions around AI design and deployment—as was the case with community advocates and many frontline workers—these questions would be skipped, and more time would be spent on discussing these participants’ desires for improved decision processes. We report findings on how agency decision-makers currently make decisions around AI, including how their decisions are shaped by complex power relations they hold with stakeholders external and internal to their agency (e.g., legal systems, frontline workers), in [217]. In this paper, we share complementary findings that provide design rationale for the *Situate AI* Guidebook’s design.

Co-Designing the Deliberation Questions

In the co-design activity, we first presented each participant with three potential scenarios: (1) Discussing *ideas* for new algorithms to improve services, (2) Deciding whether to pursue the *development* of a given algorithm design to improve services, and (3) Deciding whether to *adopt* an existing algorithm already implemented by others. We asked the participant to pick the scenario they had the most experience in or faced the most challenges for. For the scenario they selected, we asked the participant to think about what critical considerations and questions they believe should be on the table, when deliberating around these scenarios in an ideal future situation. If the participant was having a challenging time thinking of potential considerations and questions, we provided them with examples that were directly based

on challenges they had brought up during the semi-structured interview (if applicable). To help document and organize, in real-time, the considerations and questions the participant was bringing up, we shared our screen and a link to an online board on Mural, a collaborative web application where multiple users can generate and arrange sticky notes. See Figure 6.2 for an example of a blank canvas.

We asked each participant to brainstorm critical considerations and questions they would want future agencies to discuss. To avoid biasing participants, they were initially asked to openly ideate their own questions without viewing questions generated by prior participants. Following this, participants were shown existing questions, providing them an opportunity to comment upon and validate existing questions generated by other participants. As the participant openly generated ideas for questions, one of the members of our research team took post-it notes on what they were saying on the Mural board. The researcher would frequently check in with the participant, to ensure the post-its accurately represented their ideas. We also welcomed them to edit the post-its or create new ones. As they brainstormed, we asked follow-up questions to better understand how they think a given question could get answered, what makes it challenging to answer the question now, or how they are conceptualizing certain terms. For example, when a participant generated the question “How well are we involving community members in these decisions?,” we asked them to further elaborate on what this might look like in practice. This generated additional post-its, like “How well do we understand the costs, risks, and effort required of community members, if we invite them to contribute to model design decisions?” and “How are we weighting false positives and false negatives in a given algorithm, based on what type of mistake that is for the impacted community members?”

As mentioned above, after the participant generated their own questions and considerations on the blank canvas, they were shown a list of topics and example questions for additional consideration. This helped scaffold further ideation on any considerations they may have missed in their initial ideation. In the first study session, we provided an initial list of eight broad topics, informed by prior literature: (a) Overall goal for using algorithmic tool, (b) Selection of outcomes that the algorithmic tool should predict, (c) Empirical evaluations of algorithmic tool, (d) Legal and ethical considerations around use of algorithmic tool, (e) Selection of training data for algorithmic tool, (f) Selection of statistical models to fit data, (g) Long-run maintenance of algorithmic tool, and (h) Organizational policies and resources around use of algorithmic tool. We prompted the participant to discuss any new ideas the provided topical categories inspired, or any disagreements they had with the categories. Figure 6.3 shows an example of what this list looked like in later stages of the co-design process.

Between study sessions one or more researchers in our team iterated on the post-its generated during that study, to reduce redundancies and improve clarity. We then grouped the individual questions and considerations underneath the existing topical categories, while iteratively refining categories or creating new categories and subcategories as needed. The next participant was shown this updated version of the aggregated questions and topics at the end of the study.

Guidebook Reflection and Validation

Participants that contributed to later stages of the co-design process were shown an overview of the aggregated questions, a recommended deliverable for the deliberation guidebook, and a high-level outline of a proposed deliberation process. We first

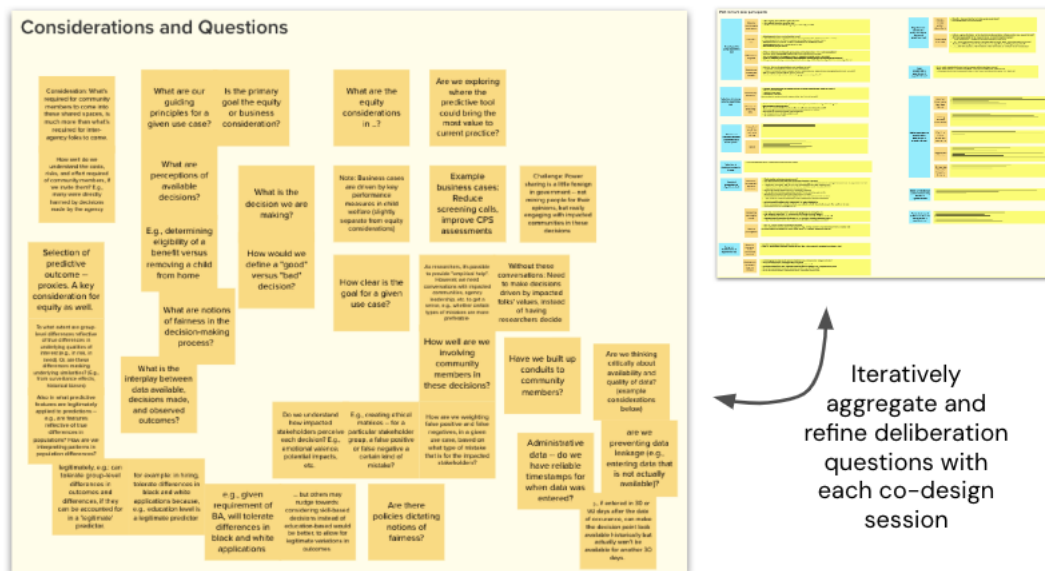


FIGURE 6.3: Screenshot of a Mural board populated with post-its after one participant's co-design activity. In our iterative co-design process, these post-its were refined by the research team then added to an aggregated list of questions that were successively grouped into higher level categories.

showed participants the aggregated questions, and asked if there were any questions that they felt were critical to include but missing. We additionally asked if they disagreed with the importance of any of the questions, or if the wording of any question was confusing in any way. We then showed the participant an overview of the deliberation process and asked for their perspectives around what they would like to change in the proposed process, to have it fit better into their existing organizational decision-making processes. To address challenges in the potential use of the protocol, we also asked participants (especially frontline workers and community advocates) about challenges they anticipate with participating in the deliberation process. We then invited participants to discuss potential adjustments to the process or alternative processes that can help address these challenges and create a safer environment for them.

6.3.3 Qualitative Analysis

The study recordings from the semi-structured interviews and co-design activities were transcribed and then qualitatively coded by two members of the research team using a reflexive thematic analysis approach [39]. We ensured that all interviews were coded by the first author, who conducted all of the interviews and, whenever applicable, another author who observed the interview. The first author coded one transcript first, then discussed the codes with other coders to align on coding granularity. Each coder prioritized coding underlying reasons why participants generated certain questions during the co-design activity, while also remaining open to capturing a broader range of potential findings. We resolved disagreements between coders through discussion.

6.4 The *Situate AI* Guidebook

The *Situate AI* Guidebook is a process to scaffold early-stage deliberations around *whether and under what conditions* to move forward with the development or deployment of a proposed AI innovation. The current version of this toolkit is intended for use within public sector agencies at various stages of maturity in their use of AI tools—from those that are just beginning to consider the use of new AI tools to those that may already have years of experience deploying AI tools. The deliberation questions are designed to be discussed across different stakeholders employed in a public sector agency, such as agency leadership, AI practitioners and analysts, program managers, and frontline workers.

In this section, we provide an overview of the *Situate AI* Guidebook as an outcome of our co-design and validation sessions. We describe the *Situate AI* Guidebook through the following sections: (Section 6.4.1) Guiding Design Principles, (Section 6.4.2) Content Design, (Section 6.4.3) Process Design, and (Section 6.4.4) Success Criteria for Use.

To provide context for key design decisions, throughout each section, we elaborate on participants' existing practices, challenges, desires, and needs for improving their decision-making process, drawing upon our thematic analysis. Where appropriate, we describe how the *Situate AI* Guidebook compares with existing responsible AI toolkits. At times, participants diverged in their desires (e.g., regarding how the decision-making process should be integrated into their agency). In some of these cases, our research team integrated these disagreements into the design of the guidebook (see Design Principle 2); in other cases, we document how these disagreements present challenges for the use of the guidebook, suggesting opportunities for future work (Section 6.4.3 and Section 7.14).

6.4.1 Guiding Design Principles

The goal of the guidebook is to scaffold public sector agency decision-making around the following question: *Should we move forward with developing or deploying a proposed AI tool? If yes, what are key considerations to plan for?* The guidebook aims to support agencies in answering this question through a deliberation-driven process supported by the following materials: (1) Question prompts to support conversations around the social (organizational, societal, and legal) and technical (data and modeling) considerations that should inform their recommendation, (2) Pointers to external resources to help guide their responses, (3) Template for a recommended deliverable to help communicate rationales and evidence for the recommendation that results from these deliberations, (4) Proposed use cases that illustrate how agencies could adopt the guidebook into their existing work processes, and (5) Success criteria to signal whether the intended outcomes of the guidebook may be relevant and useful to agencies.

In co-designing the guidebook towards this goal, we centered two core design principles:

- **(Design Principle 1) Promoting reflexive deliberation.** The question prompts (Section 6.4.2) should support stakeholders in having reflexive discussions—for example, conversations that surface their own pre-existing assumptions and beliefs about human versus AI capabilities and limitations with respect to a given task and context, or that surface relevant tacit knowledge that may be helpful to share with others. The question prompts should be designed to

avoid prompting simple yes or no responses, to ensure that responses to complex questions are not reduced to a simple compliance activity. In drawing on prior work that emphasize the role of the toolkit as one that “prompts discussion and reflection that might not otherwise take place” [279, 380], this design principles extends these notions of effective toolkits from prior literature to apply to topics of importance in public sector contexts. Prior research on public sector contexts (e.g., [182, 221, 395]) as well as findings from this study suggest that agency stakeholders’ differing backgrounds shape their assumptions and concerns around AI tools, motivating the need for a deliberative decision-making process that surface these individual differences. Throughout Section 6.4.2, we elaborate on participants’ existing challenges and desires to illustrate the importance of Design Principle 1 in their contexts.

- **(Design Principle 2) Ensuring practicality of the process.** The guidebook should be designed to support a process (Section 6.4.3) that public sector agencies can feasibly understand, adopt, and adapt as needed. If an agency already has an existing decision-making structure, or conversations related to AI design already take place, the agency should find it easy to “fit” the guidebook into their existing organizational processes and conversations. This design principle is aimed at addressing concerns raised in prior literature that existing responsible AI toolkits are often designed in isolation from the organizational contexts they intend to augment (e.g., [453]). This design principle is also motivated by our observations of the four public sector agencies in our study, which each had their own existing or planned organizational processes for developing AI tools (Section 6.4.3). Further, by co-designing the *process* the toolkit should follow (in addition to the guidebook content), we further an understanding of how organizational, labor, and power dynamics implicate the potential effectiveness of responsible AI toolkits in the public sector (Section 6.4.3).

6.4.2 Content Design: Scaffolding Reflexive Deliberation

Participants ideated critical questions that spanned four high-level topics, 12 mid-level, and 20 low-level topics. In each of the four subsections below, we briefly describe why participants were interested in the overall category of questions and provide example questions. The full set of deliberation questions for the *Situate AI* Guidebook (Ver.1) can be found in Appendix B.1.

Goals and Intended Use

This section is intended to scaffold conversations around the following broad questions:

1. Given our underlying goals and intended use case(s), is our proposed AI tool appropriate?
2. What evidence do we have to support our answer to the previous question? What additional tasks may be required in the future to help us gather more evidence and/or better understand the evidence we currently have?

Sample Questions

- Overall goal for using algorithmic tool
 - Who is going to be affected by the decision to use this hypothetical AI tool?
 - What evidence do we have suggesting that the painpoint this tool aims to solve actually exists?? What evidence do we have suggesting that technology may offer a remedy to this pain point?
 - Recall the stakeholders who are the most impacted by this hypothetical AI tool. How do we bring their voices to the table when determining goals?
 - Are there differences in the goals the agency versus community members think the tool should address? If so, what are they? If we are uncertain, what can we do to understand potential differences?
 - What biases (as a public sector agency) do we bring into this decision-making process?
- Selection of outcomes that the algorithmic tool aims to improve
 - Hypothetically, imagine that our tool does a perfect job of improving the outcome that it targets. What additional problems might this create elsewhere in the system?
- Empirical evaluations of algorithmic tool
 - Once the tool is deployed and in use, how can we evaluate how well it is working in the short-term? How can we evaluate how well it is working longer-term?
 - How can we effectively evaluate the tool from the perspective of impacted community members?
 - How might frontline workers respond to the tool? How can we better understand their underlying concerns and desires towards the tool?

The deliberation questions focus on promoting conversations that bridge reflection and understanding of the goals of the proposed AI tool, as well as how these goals will be operationalized into measurable outcomes. The 52 questions within the *Goals and Intended Use* section are divided into nine subsections: (1) Who the tool impacts and serves, (2) Intended use, (3) How agency-external stakeholders should be involved in determining goals, (4) Differences in goals between the agency and impacted community members, (5) Envisioned harms and benefits, (6) Impacts of outcome choice, (7) Measuring improvement based on outcomes, (8) Centering community needs, and (9) Worker perceptions. For the purpose of this paper, we sample one question from each topical subsection.

Several of the questions in this section are designed to help **surface underlying assumptions regarding who benefits** from the use of the tool, and to support discussion around **what evidence suggests that these assumptions are true**. These questions stem from participants' concerns around whether their AI systems are targeting areas that would bring the most benefits, and to whom these benefits apply.

For example, one participant noted that their agency had invested a lot of effort into assessing and trying to improve fairness in their algorithms. However, the participant wondered whether they should have been having conversations around larger, “more challenging” questions. For instance, they wondered whether “correcting for bias” in an algorithm within an inherently biased system is a meaningful or feasible goal. They further elaborated:

“I think there my concern often has to do with [the] unexamined belief that an algorithm is always an improvement. [...] I think [questions on broader goals and benefits are] more challenging and that people [who] are running the system may not always see [...] Personally, I think there’s a lot of stuff that can be done with machine learning that doesn’t have to [target] decision-making at the participant level. [...] But those are the kinds of questions the immediate focus [is] on. ‘Oh, we’re going to use this to make decisions at critical points in programs.’ Those are things that to me still need to be discussed. And it may be that those conversations are happening at tables that I’m just not at.” (A02)

Other participants expressed concerns for how frontline workers in their agency—the majority of who are currently not involved in early-stage conversations around the goals of the AI tool—may be **misunderstanding the intended uses and capabilities of their AI tools**. For example, one participant described that frontline workers may be concerned that the AI tools will displace them, even though their agency doesn’t intend to use them to automate workers’ jobs. They described:

“There’s almost like a mystique around machine learning algorithms, like there’s some amazing thing that is all knowing and all seen, and therefore can predict all these different things. [...] helping people [...] understand] what it’s able to do and not able to do, I think, is something we’ve struggled with” (A04).

Other questions are intended to help **forefront considerations around what additional planning and resources may be needed**, in order to adequately complete a related task in the future. For example, workers within agencies often described that involving community members in their AI design and evaluation process can be challenging, given the current lack of infrastructure to support such collaborations. However, community advocates described how involving community members is often an after-thought. One community advocate described the importance of being intentional and proactive in community engagement practices, because

“it’s easy to let that be something that gets back burning, like throughout the process to just have that be something we’ll get to, and then we end up in that feedback loop where the feedback is provided but the tool is already created” (C2).

Questions in this section help promote earlier reflection and planning on how community members could be involved, so that they could conduct appropriate empirical evaluations regarding their perceptions of the AI tool.

Societal and Legal Considerations

This section is intended to scaffold conversations around the following broad questions:

1. Given the societal, ethical, and legal considerations and envisioned impacts associated with the use of AI tools for our stated goals, is our AI tool appropriate?
2. What evidence do we have to support our answer to the previous question? What additional tasks may be required in the future to help us gather more evidence and/or better understand the evidence we currently have?

Sample Questions

- Legal considerations around the use of algorithmic tool
 - Do the people impacted by the tool have the power or ability to take legal recourse?
- Ethical and fairness considerations around the use of algorithmic tool
 - Are there differences in the goals the agency versus community members think the tool should address? If so, what are they? If you are uncertain, what are your plans for understanding potential differences?
 - Can we agree on a definition of fairness and equity in this context? What would it look like if the desired state is achieved?
 - Are fairness and equity definitions and operationalizations adequately context-specific? (For example, in the child welfare domain: children with similar profiles receive similar predictions irrespective of race?)
 - Do we understand the negative impacts of the decision made across sensitive demographic groups?
- Social and historical context surrounding the use of algorithmic tool
 - Have we recognized and tried to adjust for implicit biases and discrimination inherent in these social systems that might get embedded into the algorithm?
 - How might we clearly communicate the limitations and historical context of the data to community members?

Overall, the goal of this section is to help promote a systematic, deeper conversation on the various dimensions of social and ethical concerns relevant to the design of an AI tool. The 38 questions within the *Societal and Legal Considerations* section are divided into seven subsections: (1) Legal considerations around the use of the algorithmic tool, (2) Impacted community member needs, (3) Involving impacted communities, (4) Clarity of ethics goals and definitions, (5) Operationalization of ethics goals, (6) Envisioning potential negative impacts, and (7) Social and historical context surrounding the use of the algorithmic tool. Again, for the purpose of the paper, we sample one question per topical subsection.

Participants shared that they did not currently have **structured opportunities**

to proactively discuss social and ethical considerations surrounding AI tool design. While participants described that their teams spent a lot of time working on related data- and model-specific fairness tasks (e.g., using bias correction methods to improve the fairness of their AI tool), several participants noted a **desire to discuss normative concerns regarding the design of an AI tool that could only be addressed in earlier problem formulation stages**. Moreover, participants' past experiences illustrated an opportunity to better support cross-stakeholder communications around the ethical considerations that should aid AI design, by equipping teams with a **shared knowledge base and vocabulary** for ethical concerns. For instance, one participant described how a leadership team tasked them with creating a predictive algorithm to assist decisions about fraud investigation. The participant's team tried to "get them away from this" because the task was technically infeasible (producing high false positive rates) and ethically risky the cost of errors is high, given that decisions to investigate are highly intrusive to the individual. This section's questions intend to support agency stakeholders in forming a more complete understanding of the different ethical factors that could make a proposed AI tool design "appropriate" or "inappropriate."

We note that the guidebook does not exclusively surface societal and ethical considerations in this section; the prevalence of relevant questions included in the other three topical sections (Goals and Intended Use, Data and Modeling Constraints, Organizational Governance) reflect how social and ethical considerations are intertwined with all facets of a proposed AI tool.

Data and Modeling Constraints

This section is intended to scaffold conversations around the following broad questions:

1. Given the availability and condition of existing data sources, and our intended modeling approach, is our proposed AI tool appropriate?
2. What evidence do we have to support our answer to the previous question? What additional tasks may be required in the future to help us gather more evidence and/or better understand the evidence we currently have?

Sample Questions

- Understanding data quality
 - Has the definition of the data changed over time? (E.g., in child welfare, has reunification always meant to reunify with the parent?)
- Process of preparing data
 - How are we preprocessing the data?
 - Who should be involved in making decisions around whether to include or exclude certain data points or features? Do we have plans for involving those people?
- Model selection
 - Is our model appropriate given the available data? Why or why not?

This section intends to forefront conversations around data and technical work that may be critical to have earlier on. The 18 questions within the *Data and Modeling Constraints* section are divided into seven subsections: (1) Understanding data quality, (2) Process of preparing data, (3) Selection of statistical models to fit data. For the purpose of the paper, we provide a subsample of questions under each topical subsection.

Participants who had experience developing AI tools often underscored the importance of ensuring that they had the computing resources and data needed to develop their proposed AI tool. For example, they described the importance of forming a context-specific understanding of the data labels that may be challenging to identify without relevant domain knowledge (e.g., whether certain labels like “reunification” have changed definitions over time). Others described the importance of deliberating who should be involved in data inclusion and exclusion decisions when they are cleaning their data.

Organizational Governance Factors

This section is intended to scaffold conversations around the following broad questions:

1. Given our plans for ensuring longer-term technical maintenance and policy-oriented governance, do we have adequate post-deployment support for our proposed AI tool?
2. What evidence do we have to support our answer to the previous question? What additional tasks may be required in the future to help us gather more evidence and/or better understand the evidence we currently have?

Sample Questions

- Long-run maintenance of algorithmic tool
 - Do we expect there will be shifts in performance metrics over time? If so, why? What are our plans for identifying and mitigating those shifts?
 - Do we have the mechanisms to monitor whether the tool is having unintended consequences?
- Organizational policies and resources around the use of algorithmic tool
 - Is there training for frontline workers who will be asked to use the tool? What evidence suggests that this training is adequate?
 - Imagine that we could assemble the “ideal team” to monitor and govern the tool after it is deployed: What are the characteristics of this ideal team?
 - * Who is the *actual* team that will monitor and govern the tool after it is deployed?
 - * Given the gaps between the “ideal team” and the actual team we expect to have: What risks to post-deployment monitoring and governance can we anticipate? How might we mitigate these risks?

- Internal political considerations around the use of algorithmic tool
 - How well do we understand system administrators’ and leadership’s perspectives around the use of this tool?
 - How well do staff and leadership understand ‘why’ the tool could bring value?

The 24 questions within the *Organizational Governance Factors* section are divided into five subsections: (1) Measuring changes in model performance over time, (2) Mechanisms to identify long-term changes in model performance, (3) Policies around worker interactions with the AI tool, (4) Governance structures around the AI tool, and (5) Internal political considerations around the use of the AI tool. As with prior sections, we include in this paper a sample of questions across these topical subsections.

Similar to considerations around the *Social and Legal Considerations* of AI design (Section 6.4.2), participants often described encountering challenges when attempting to meet organizational governance-related needs of the AI tool, like maintaining their AI tool over time, ensuring workers are adequately trained, or communicating the goals and capabilities of the AI tool to agency leadership. Participants highlighted that many of these challenges arise because such considerations are discussed in an ad-hoc manner, too late in the AI development process. Given that several of these needs may require longer-term planning and preparation (e.g., gathering resources of model maintenance), public sector agencies may be better equipped in meeting these governance needs if they were discussed in early stages of model design (rather than after an AI tool has already been developed). For example, participants described how they currently lack domain experts that could help maintain and improve their model post-deployment—a gap in their AI development process that they felt was critically important to address. While agencies currently discuss maintenance-related concerns at the deployment stage, this may not allow the agencies enough time to deliberate who should be involved in maintenance, or how to allocate additional roles for a maintenance team.

6.4.3 Process Design: Designing for Practicality and Adaptability

The overall goal of the *Situate AI Guidebook* is to help public sector agencies make more informed, deliberative decisions about whether and how to move forward with implementing a proposed AI tool. Prior literature studying existing responsible AI toolkits have started to surface concerns around how such toolkits may be used inappropriately or not used at all in practice, due to misalignments with the organizational contexts they are designed to support [279, 453]. In this section, we describe findings related to the broader deliberation *process* that participants envisioned the deliberation questions (Section 6.4.2) could be used to support.

Below, we first present our proposed use case for the *Situate AI Guidebook*, along with an example instantiation of the use case and an explanation of how participants’ existing practices informed this use case. We then discuss participants’ desires for alternative use cases and processes around deliberation. Participants across agencies and roles expressed interest in using the questions in a few different ways, based on their concerns around cross-stakeholder power dynamics and desires to enable deliberation practices that align with their organizational values [337, 453].

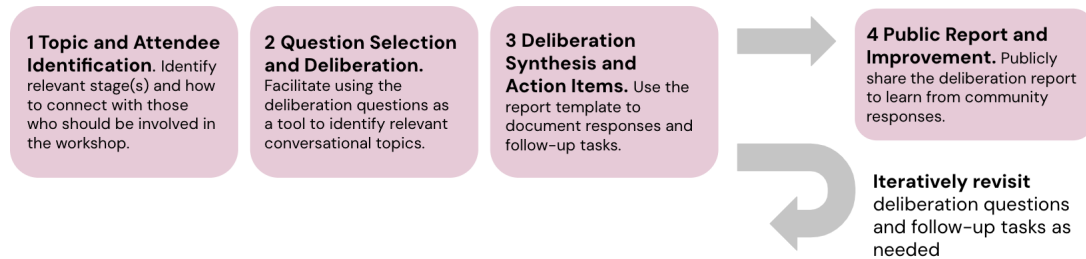


FIGURE 6.4: A high-level overview of the main stages involved in the proposed use case for the *Situate AI* Guidebook, intended to support structured and iterative deliberations within a public sector agency.

Given participants' interests in adapting the guidebook to different use cases, a key component of the *Situate AI* Guidebook is that it is designed to allow users to select which topics and questions they would like to focus on: The deliberation questions are categorized and grouped into modular components; and users have the flexibility to select from a large set of deliberation questions within each component to identify a subset that is most relevant to their use case.

Proposed Use Case: Using the Guidebook to Support Structured and Iterative Deliberations

Participants envisioned that the guidebook could be effectively used to support structured, iterative deliberation through formal workshops between members of their agency. In this section, we elaborate on one possible way this use case can be instantiated into an overall deliberation process, then discuss how this compares to participants' existing practices within their agency. We provide an example of one possible implementation of a formal deliberation process, using the guidebook.

Example instantiation of proposed deliberation process This process involves a four-stage phase, where the public sector agency would first appoint a facilitator(s) to help organize the overall decision-making process:

Stage 1: Topic and Attendee Identification. The facilitator will identify which of the four guidebook topics, if not all, they are interested in convening a deliberation workshop on. Based on broad guidance provided in the guidebook, the facilitator will then identify the stakeholders that should be included in the deliberation workshop based on the selected topics. For example, the Societal and Legal Considerations section is designed to be used by a more diverse range of stakeholders (e.g., AI practitioners, frontline workers, community members, legal experts) compared to the Data and Modeling Constraints section (e.g., AI practitioners only).

Stage 2: Question Selection and Deliberation. After finding a shared time for the deliberation workshop, the facilitator should share the goals and topics of deliberation (included in the guidebook) with the group. The guidebook includes a large number of questions for each topic of deliberation. For example, the *Goals and Intended Use* section alone has 52 questions. To ensure that the questions can be feasibly discussed within the allocated time, we highlight 1-2 recommended questions per major subsection, resulting in a smaller number of questions (19 questions for the *Goals and Intended Use* section). The remaining questions are also available in the guidebook as "optional questions." We recommend that, at the start of the workshop, the facilitator provides the attendees with the opportunity to identify any

questions from the “optional questions” category they would like to additionally or alternatively discuss in the workshop. As the attendees are discussing each question, the facilitator should take note of their responses and points of disagreement. If there are disagreements that are challenging to resolve in response to a question, the facilitator should help the group identify action items to help gather more information or perspectives and plan to revisit the question at a later time. If the group finds they currently lack the resources or knowledge to fully address a question, the facilitator should also make note of this and plan to revisit the question at a later time.

Stage 3: Deliberation Synthesis and Action Items. After the deliberation workshop, the facilitator should summarize the discussions and outcomes into the deliberation report template which we include in the guidebook. The template includes the following questions: (1) *What is your recommendation?* (2) *Please list core reasons for your recommendation, based on the deliberation workshop,* (3) *Are there any follow-up tasks you must complete, in order to fully support this recommendation? If yes, please write the task(s) and plan(s) for completing it,* and (4) *What core counter-arguments against this recommendation arose during the deliberation workshop? Please describe each counter-argument, including how you addressed or plan to address each one.* Based on the report responses, the facilitator should continue to iteratively revisit the deliberation questions, organizing additional deliberation workshops with the attendees as needed, as they complete the follow-up tasks included in the report.

Stage 4: Public Report and Improvement. The facilitator should work with their agency to publicly share the deliberation report with agency-external stakeholders, including impacted community members and related organizations. To promote conversations and bidirectional learning between community members and agency-internal stakeholders, the agency should hold public convenings and host an online commenting forum for any individuals who would feel more comfortable contributing anonymously online. The agency should then synthesize the themes that emerged from the conversations, identify action items to address any concerns, and share these results of the community conversations with the public. In the guidebook, we plan to include links to existing community review efforts to provide examples of what this interaction could look like. However, we note that effectively completing this step requires additional research and resource creation (which we discuss in the Discussion Section 7.14).

How participants’ existing organizational practices informed the proposed process design Reflected in the process above, participants raised several important considerations to ensure the process is practical and meaningful to their agency. For example, participants in agencies that are actively developing new AI tools described that there are already AI design and development processes in place that support focused discussions on improving, for example, the algorithmic fairness of their AI tool. These participants did not want the Situate AI Guidebook to replace these conversations and work sessions. Instead, they desired a process that could **augment their existing processes**. For instance, participants noted that having these deliberation workshops earlier on, before they developed or analyzed any AI model, can help promote reflexive conversations about what it means to do the work of AI fairness and what important considerations that should aid this work (e.g., whether there is a definition of “fairness” that agency workers agree on). Relatedly, as reflected in the process description, several participants described the importance of **revisiting these deliberation questions iteratively**, throughout the AI development and deployment process (rather than only discussing these at the early

ideation stages of a development process). For instance, participants described that their understanding of some of the question responses (e.g., the intended outcomes that the AI tool should help achieve) may evolve with time as the AI tool development evolves (e.g., depending on what is technically possible, given data constraints or prediction errors). Participants noted that designing the guidebook to center an iterative process would help ensure the conversations complement their existing AI development process, which is also iterative in nature. We originally designed the process so that the deliberation workshop attendees would be required to come to a consensus on the final recommendation before moving forward. However, the participants we spoke with emphasized that this may be impractical and unnecessary based on their end goals for the guidebook. We elaborate on this point in Section 6.4.4 when we discuss the guidebook's Success Criteria.

We discuss limitations and future work related to this proposed use case in the Discussion, drawing on prior literature suggesting ways in which research-based conversational tools may not be adopted in practice or may be used in inappropriate ways.

Empowering Participation: Accounting for Organizational Power Dynamics.

Findings from this study strongly suggest that frontline workers—those who would be asked to use the AI tool once deployed—are interested in engaging in early-stage deliberations around what the AI tool should be designed to assist. Prior literature also suggests that ensuring agency leadership and AI developers understand frontline workers' needs and challenges is critical to ensuring that the "right" AI tools are being developed (e.g., [217, 221, 361, 459]). However, effectively supporting conversations between roles with prominent and knowledge differentials remains a challenging task [182]. For the current version of the *Situate AI* Guidebook, we begin accounting for this challenge by editing the language used in some of the guidebook sections to ensure it is understandable to those without prior knowledge on technology. We additionally asked frontline workers about their desires for how they would want to participate in the deliberation process, to ensure they feel safe to share any concerns. In this section, we discuss these findings. However, we note that future work is needed to ensure that the *Situate AI* Guidebook (Ver.1.0) adequately accounts for organizational power dynamics. In the Discussion, we discuss implications for complementary policy interventions that may be needed for an agency to effectively facilitate a multi-stakeholder deliberation process.

Frontline workers' preferred processes for participating in multi-stakeholder deliberations Frontline workers in our study had a range of perspectives around how best to involve them and their colleagues in the deliberation process. One frontline worker suggested that their agency should require all frontline workers to attend the deliberation workshops. They described that, without making participation in these discussions mandatory, frontline workers may opt to skip meetings given their busy schedules. The participant further expressed concerns that, if participation was on a voluntary basis, frontline workers who join may be harmfully self-selective: "... particularly for people with marginalized identities, it's important for them to be a part of these spaces and voice their concerns. I think that if it wasn't mandatory, it might be, you know, [a] self-selecting group" (F7). This frontline worker, along with another worker, also described how adjustments to the proposed process could help frontline workers feel more comfortable raising concerns. For example, the participants expressed that, for multi-stakeholder conversations, some

frontline workers may feel more comfortable having a separate frontline worker-only deliberation workshop, synthesizing and formalizing their perspectives, then going to a group meeting with other agency stakeholders to present their perspectives. As the participant described:

“A lot of social workers are very non-confrontational [...] we are our clients’ best advocates but not for ourselves. And so I definitely do think that people might be more comfortable, you know, having their own sort of peer group discussion or colleague group discussion. And then that being you know, sort of the concerns being written down and formalized, and that being taken rather than a more informal sort of like anyone who has concerns just raise their hand and say their piece. I feel like that might be a bit daunting for some people” (F7).

Alternative use case: Using the guidebook to empower everyday conversations

Complementing frontline workers’ desires for engagement, participants from agencies that were not yet developing new AI tools described a different use case for the Situate AI guidebook. These participants envisioned that the guidebook could be used by teams to support everyday conversations, with the aim of proactively avoiding pitfalls in AI project ideation and selection. For instance, one participant described that they wanted the guidebook to be used more casually, by everyone in the agency, to help all staff members feel empowered to “be able to do a little more of the innovation” (L7). The participant described that, even if they get stuck or need help, “it would be awesome for them to have a library of resources that they can look at” to help them get started. The participant further described that workers in their agency should “have the flexibility to structure the deliberation workbook to their needs,” for example, deciding which questions to discuss, how much time to take in discussing the questions, and who to talk with. By having a guidebook that empowers any staff member to discuss topics around AI, the participant hoped that these deliberations could have rippling effects on their agency’s overall culture:

“Maybe we’re trying to get from [...] ‘let’s do this big project right with the right leaders and things in the room’ to ‘how do we create a culture of improvement’ [...] Not just how do we do a technology project the right way, but actually can it have a broader impact on culture. This is how we do anything. It’s always with this batch of questions in mind and thinking about how we can be people around problem solving” (L7).

Desires to expand participation to community members for future versions of the guidebook

Several participants across agencies and community advocacy groups were interested in involving community members in the deliberation workshops supported by the guidebook. Workers within the agency wanted guidance on how to do this effectively. In our conversations with community advocates, we probed on how they would like to be involved in deliberations around the Situate AI guidebook. They described the importance of compensating community members for their time, providing multiple channels for communication (e.g., online forums and in-person meetings), and following up with the outcomes of the conversations: “that happens a lot, you know—agencies are like ‘oh, we engage with the community, and we brought them into the space with us.’ But then there’s no follow up or follow through from those conversations. And that’s been a historical thing” (C2).

Importantly, we note that the guidebook, in its current form, is designed to support conversations across workers within a public sector agency. It is not designed to directly support conversations between agency-internal workers and agency-external stakeholders (e.g., community members). In the Discussion section, we discuss opportunities to expand participation through design improvements.

6.4.4 Success Criteria

What outcomes do public sector agencies and impacted community members consider “meaningful,” when assessing the effectiveness of the *Situate AI* Guidebook? What are their underlying theories of change around how their public sector agency could move towards more responsible early-stage AI design practices, and how do they envision the *Situate AI* Guidebook can help them progress towards that path? Overall, through the guidebook, participants wanted to form an understanding of the disagreements and tensions across agency workers felt most strongly about, to help position themselves to better address these disagreements through changes to the problem formulation or design of a proposed AI tool. Importantly, as indicated by the process design in Section 6.4.3, participants described that the goal of the *Situate AI* Guidebook should *not* be to *resolve* these tensions and disagreements across individuals. Participants described that this is an infeasible task, given underlying differences in values and goals across agency stakeholders. In this section, we elaborate on four success criteria of the guidebook. These success criteria intend to help communicate the intended goals and boundaries of the guidebook, and including how they are informed by participants’ own notions of success for deliberations around the design of AI tools.

Make it easier for different agency stakeholders to communicate with each other about AI design, evaluation, and governance considerations.

Many of the challenges that participants in our study described could be addressed if better communication channels existed between different agency stakeholders—including amongst AI developers, agency leadership, and frontline workers. This challenge is also well-documented in prior literature studying public sector AI decision-making [182, 217, 222, 361]. For instance, in current practice, frontline workers are often not meaningfully involved in early-stage deliberations around AI design and adoption. As a result, agency leadership and AI developers have interpreted workers’ concerns around AI as a signal for not understanding what the AI tool does. Involving frontline workers in these earlier discussions can both help more proactively inform workers of AI capabilities and empower them to engage in constructive conversations that would improve the design of AI tools.

Bring context-specific needs for resources to the forefront of AI project selection conversations.

While participants often knew which resources they needed to successfully implement a given AI tool, their past challenges sometimes surfaced a missed opportunity to identify these needs at an earlier stage of their AI design process. Moreover, related to the previous success criterion, our conversations with the participants surfaced ways in which having agency workers with different roles and perspectives engaged in these early-stage deliberations can strengthen their ability to anticipate the potential impacts of AI design decisions. For example, frontline workers voiced

that AI tools they had used in the past were designed in ways that conflicted with their existing decision-making policies; other workers described that AI deployments may add additional labor to their day-to-day tasks, given they may be asked to more diligently collect data. In current AI development processes, where front-line workers may only be meaningfully engaged in AI implementation or piloting stages, mitigating these negative impacts may require more substantive tasks like redesigning the AI tool. Participants further described that these resource-related needs were highly context-specific. For instance, when discussing the importance of anticipating how their AI tool may impact community members, one participant recalled how even the definition of “community” may differ across agencies and AI tools: “Because we would always say, ‘we’re doing stuff [where] the community is informing us.’ And then we realized, ‘oh wait, it wasn’t necessarily the group of people who were impacted by [our decisions]’” (L7).

Make social and ethical considerations a first order priority in conversations around whether to move forward with an AI tool idea.

As described in Section 6.4.2 and 6.4.2, participants described their past assessments around whether an AI tool was appropriate to implement centered algorithmic considerations—whether that be the quality of their training data or outcomes of algorithmic fairness or accuracy metrics. While these considerations are critically important, others also discussed a desire to rigorously deliberate the underlying values embedded in design decisions, and the social and ethical impacts of a proposed AI tool. This echoes concerns from prior literature, discussing how technical design decisions often include hidden policy decisions and value judgements [138, 395]. Prior work also suggests the importance of these considerations, noting that existing AI ethics toolkits have largely framed the work of ethics as “technical work” [453].

Make “fitting” an AI tool into a workplace a design problem, rather than an implementation problem.

This success criterion intends to avoid practices where the AI tool idea is conceived before fully understanding context-specific practices and needs, and in turn, creating AI tools that frontline workers must then attempt to “fit” into their existing workflow. This tendency to treat “fitting” AI tools as an implementation problem, and its negative impacts on workers’ ability to improve their existing decision-making practices, is also well documented in prior literature [459, 463]. Indeed, participants—including both AI developers and frontline workers—described that they wished they could have had better conversations, early on, to understand what the actual goal of the tool they were building should be. As one AI developer described, recalling a past experience in their team where leadership had asked them to create an AI tool: “It was kind of hard to get a sense of what the actual issue was that was being asked to be solved. It sounds kind of a lot like, ‘Here’s a bunch of different potential places an algorithm might fit in’” (A04). Ultimately, they were asked to create a predictive model to “to find fraud where there wasn’t already suspicion of fraud.” However, the AI developer described feeling leadership had proposed the idea as “this cool thing we could do” but, in reality, realizing that creating such a tool would create more problems downstream in the system (in this case, it would create too many referrals to be able to investigate). By promoting early-stage, structured deliberations around critical topics related to AI tools, public sector agencies could be supported in identifying higher-value, lower-risk opportunities to innovate with AI systems.

6.5 Discussion

Public sector agencies in the U.S. are increasingly exploring how new AI tools can assist or automate services in child welfare, homelessness housing, healthcare, and policing, among other domains [73, 143, 310, 361, 459]. In the U.S., these public services have historically been characterized by racial inequity, procedural injustice, and distrust from the impacted communities [42, 119, 395]. While agencies have rapidly begun to deploy AI tools to improve their services, ensuring responsible development has proven to be an immense challenge. In the past decade, such AI tools have often failed to serve the needs of the communities that agencies are expected to serve [21, 119, 138, 163, 164]. A growing body of literature has recognized that many downstream harms resulting from AI tools can be traced back to decisions made during the earliest problem formulation and ideation stages of the AI lifecycle. Yet, there are few, if any, effective resources for public sector agencies in making more deliberate decisions regarding *whether* a given AI proposal should be developed in the first place.

Through iterative co-design sessions with 32 individuals (agency leaders, AI developers, frontline decision-makers, and community advocates) across four public sector agencies and three community advocacy groups, we created the *Situate AI Guidebook* (Ver.1.0). The guidebook, designed for public sector agency workers, scaffolds the process for early-stage deliberations around *whether and under what conditions* to move forward with the implementation or adoption of a new AI tool or idea. To support this process, the guidebook presents a set of 132 deliberation questions—which participants indicated are critical to consider yet are often overlooked today—spanning both social (organizational, societal, and legal) and technical (data and modeling) considerations around AI; along with guidance on the overall deliberative decision-making steps and success criteria for use. In this section, we discuss the design decisions we made in creating the guidebook, along with limitations and opportunities for future work. For each section of this discussion, we begin by summarizing relevant portions of the findings. Then, we elaborate on limitations and future opportunities to improve upon the existing guidebook.

6.5.1 Overcoming Low Adoption Rates for Responsible AI Toolkits in the Public Sector

As the research community continues to innovate new Responsible AI toolkits, recent literature has raised concerns regarding the practical efficacy of such toolkits. Prior work has found that the majority of AI ethics toolkits fail to account for the relevant organizational context, hindering their usability (e.g., overlooking guidance on *how* different stakeholders should be engaged) and effectiveness (e.g., focusing on the technical but neglecting the social aspects of AI ethics work) [453]. Public sector decision-making around service allocation is often shaped by resource and staffing shortages, and require balancing tradeoffs to meet the competing needs of a range of stakeholders (e.g., impacted community members, policymakers and regulators, politicians) [439]. Moreover, AI tools in the public sector often target socially high-stakes decisions (e.g., whether to screen in a family for child maltreatment investigation, or provide an individual with a credit loan), that have disproportionately negatively impacted the lives of historically marginalized communities. Prior work has shed light on the downstream impacts that public sector AI systems have had (e.g., [42, 346]), along with challenges to ensuring their responsible design and use (e.g., [221, 361, 439]). Through our study, we demonstrated how collaborating

with public sector agencies and community members to co-design a responsible AI toolkit—including its process and content design—can help surface and account for some of these challenges. That said, future research is needed to understand how effective the toolkit is in practice, and to surface other challenges that can only be observed through actual use (rather than through our co-design and interview study format). In the following subsections, we briefly discuss related findings and opportunities to improve the contextual design and use of the *Situate AI Guidebook* for public sector settings.

Designing for more inclusive forms of worker participation.

While AI tools for public sector contexts implicate a range of different agency-internal stakeholders, these agency workers—from agency leaders to frontline workers—often operate in silos, separated by power imbalances and knowledge differentials. We found that participants desired a range of participation structures to account for these differences. For example, some frontline workers wanted to first gather amongst others with similar occupations to prepare for the deliberation workshop, and then send in one frontline worker to attend the workshop and represent their perspectives. On the other hand, other participants suggested that there should be an organizational policy that required all frontline workers to attend the deliberation workshops alongside the AI developers and agency leaders. Future work is needed to understand the broader range of solutions that could best address these differences in workers' preferences. For example, future work could pilot different processes, where agency workers are grouped in deliberation workshops in specified configurations depending on their role and background. Through observations and retrospective interviews of these configurations, we could better understand whether having a set of deliberation questions alone is adequate to prompt meaningful conversations. Future work could additionally explore how additional resources and tools could be used alongside the deliberation toolkit, in order to effectively scaffold conversations around the appropriateness of AI design ideas. This direction would be especially critical to pursue, in order to ensure that the deliberation toolkit is accessible to those who may not have had prior exposure to AI technologies.

Incentivizing and governing responsible use

Ensuring responsible use and adoption of the *Situate AI Guidebook* may require complementary efforts from governing bodies. For example, while the U.S. does not currently require public sector agencies to document early-stage deliberations around AI, having similar forms of external forces that incentivize agencies to engage in early-stage deliberation may help ensure that the deliberation toolkit is used effectively. One way to incentive public sector agencies may be to clearly communicate how the toolkit aligns with and complements existing voluntary guidelines, such as those in the NIST AI Risk Management Framework (RMF) [308]. While the NIST RMF and NIST RMF Playbook [307] both focus on providing higher level guidance on steps to follow for responsible AI design, research-based co-designed toolkits like the *Situate AI Guidebook* can help bridge gaps between their proposed policy guidance and real-world practice. In future work, we plan to map the guidebook to the four functions captured in the AI RMF Core: Govern, Map, Measure, and Manage. For example, each question or category of questions could be assigned one or more of the AI RMF Core functions.

In future work, we plan to explore with public sector agencies community advocates, and other stakeholders how new policy and organizational interventions can support them in using the *Situate AI* Guidebook. The public sector agencies in our study, including the frontline workers, expressed interest in exploring how to use the guidebook in practice through pilots.

6.5.2 Expanding the *Situate AI* Guidebook to Engage Community Members

In public sector contexts, there is often a greater expectation that decisions center the needs of the community, including by being transparent to and engaging with the community during the decision-making process. In our study, participants expressed a desire for guidance on how to engage with community members in discussing complex topics around AI design. While the deliberation protocol is not currently designed to support such conversations, the current version poses questions that suggest follow-up tasks involving conversations with community members. For example, the question “Are we assessing the tool from the perspective of impacted community members? What evidence do we have to suggest that we are genuinely understanding their concerns and desires?” suggests that the agency should talk with impacted community members to understand their perspectives—a task that would require additional guidance and resources to complete successfully. Agency workers acknowledge they are often pushed to involve community members in their AI design work but without actionable guidance on how to do so effectively. Participants suggested linking existing relevant resources from the guidebook, to assist agencies in this regard. Moreover, community advocates in our study expressed interest in engaging in the deliberation workshops themselves.

Future work should explore ways to improve the design, structure, and process of the deliberation guidebook so that it is well-equipped to support conversations between agency-internal workers and agency-external community representatives. For example, to help bridge a shared vocabulary about AI between agency workers and community representatives, future work could begin by integrating existing resources and guidance from publicly available guides like *A People’s Guide to Tech* [330]. It is possible that the specific questions and topics this deliberation guidebook addresses requires additional scaffolding and support. Future work should explore ways to provide this support through continued collaborations with community advocacy groups. Community advocates in our study additionally expressed interest in having both the option to attend in-person workshops and to participate anonymously online. Future work could explore ways to support more democratic forms of participation online using social computing platforms (e.g., [241]) intended to facilitate and analyze deliberation about specific topics around AI.

6.5.3 Exploring how the *Situate AI* Guidebook Can Support Deliberation in Non-Public Sector Contexts

While the *Situate AI* Guidebook was originally designed for high-stakes public sector decision-making domains, there is an opportunity to adapt it to meet the needs of other AI use cases. Private and public sector settings share many organizational challenges (e.g., communication barriers across teams and occupations) and development tendencies (e.g., targeting problem spaces that AI capabilities may be ill-suited towards), that could implicate the effective design of responsible AI toolkits. Moreover, by designing for a setting with relatively high expectations and standards

for responsible design (i.e., *high stakes* AI applications in the *public sector*), the *Situate AI* Guidebook sets a high bar for the kinds of questions and processes that should be followed to responsibly evaluate early-stage AI design concepts elsewhere. For this reasons, we expect that the guidebook may be (at least partially) applicable to other AI deployment contexts, including certain high risk applications in industry (e.g., healthcare, credit lending).

Indeed, many of the questions that the participants generated are relevant to AI deployment contexts beyond the public sector. Most deliberation questions target core issues relevant to all AI deployments (i.e., around the goals, ethical implications, technical constraints, and governance practices surrounding an AI deployment). Besides the deliberation questions, the design principles underlying the guidebook may also help make the guidebook useful for non-public sector contexts. Because the public sector agencies we partnered with differed in their organizational practices (e.g., who is involved in decision-making around AI) and priorities (e.g., types of services provided), we intentionally designed the guidebook to allow for flexible adoption and personalization. For instance, in designing towards Design Principle 2, we categorized the questions into modular topics and subtopics that can be selected and combined for a given deliberation workshop. There is an opportunity for future work to expand the set of labels attached to the deliberation questions. For example, participants described an interest in future versions of the toolkit that categorized questions based on the type of technology (e.g., generative AI, predictive analytics), or the type of deliberation (e.g., individual assumption-checking, knowledge-sharing, future task identification) that the question is intended to support. Future work could similarly aim to understand the types of questions that are the most critical for certain AI deployment contexts (e.g., public sector social work, private sector healthcare, etc.).

6.6 Conclusion

As public sector agencies in the U.S. increasingly turn to AI tools to increase the efficiency of their services, it becomes critical to ensure these tools are designed responsibly. While much research and development efforts have been dedicated to better scaffolding responsible AI development and evaluation practices, real-world AI failures often point to fundamental problems in the problem formulation of an AI tool—problems that should be addressed before proceeding with any decision to develop an AI tool. Yet, we currently lack effective processes to support such early-stage, deliberate decision-making in the public sector. This paper introduces the *Situate AI* Guidebook: the first toolkit that is *co-designed* with public sector agencies and community advocacy groups to scaffold *early-stage deliberations* regarding *whether or not* to move forward with the development of an AI design concept. Through co-design sessions conducted over the course of 8 months, participants generated 132 questions which we organized under four high-level categories including (1) *goals and intended use*, (2) *social and legal considerations* of a proposed AI tool, (3) *data and modeling constraints*, and (4) *organizational governance factors*. In this paper, we elaborate on how participants' practices, challenges, and concerns shaped the *Situate AI* Guidebook's guiding design principles, the deliberation questions they believed were critical for early-stage decision-making, the overall organizational and team decision-making process the guidebook should scaffold, and the success criteria used to assess the effectiveness of the guidebook. We additionally discuss opportunities for future work to improve the design and implementation of the *Situate AI*

Guidebook, including via continued partnership with public sector agencies in our study, who plan to pilot how the guidebook can be used in their agency.

Chapter 7

Improving AI Measurement

These chapters are taken from the following papers:

Anna Kawakami, Haiyi Zhu, and Kenneth Holstein. “AI Measurement as a Collaborative Design Practice: A Multidisciplinary Synthesis and Future Directions.” Under review.

Anna Kawakami, Chloe Zhao, Renee Shelby, Fernando Diaz, Haiyi Zhu, and Kenneth Holstein. “‘I want to be pushed, I want to grow’: Enabling social workers to design evaluations of LLM augmentation in their work.” In *Proceedings of the 2026 AAAI/ACM Conference on AI, Ethics, and Society*.

Improving AI Measurement, Part 1

Framework for AI Measurement as a Collaborative Design Practice

This chapter is taken from the following paper:

Anna Kawakami, Haiyi Zhu, and Kenneth Holstein. “AI Measurement as a Collaborative Design Practice: A Multidisciplinary Synthesis and Future Directions.” Under review.

Abstract. A growing body of research has called for more stakeholder collaborative approaches to AI design and evaluation, where impacted stakeholders (with expertise in the specific real-world contexts that AI systems are deployed) shape the design and evaluation of AI systems. However, much of this research has overlooked a critical opportunity to engage stakeholders in shaping the underlying *measurement assumptions* behind AI systems—made when translating from abstract, latent concepts (e.g., “fairness”, “toxicity”) into practical approaches for measuring those concepts, such as metrics, benchmarks, or LLM-based classifiers. In this paper, we aim to guide nascent research that calls for more collaborative approaches to AI measurement by learning from past work across disciplines that has engaged stakeholders in measurement design, both within and beyond the context of AI. Based on a multidisciplinary literature review, we synthesize a methodological framework of key dimensions that mediate stakeholder participation in measurement design. Using this framework as a lens, we analyze existing approaches and identify opportunities to inform the design of new approaches. We discuss how the framework can be applied to both assess and improve upon future approaches to engaging stakeholders in AI measurement design.

7.1 Introduction

As AI systems are rapidly deployed across a range of real-world contexts, a growing body of research has described how downstream harms can be attributed to upstream decisions around the goals, design, evaluation, and governance of AI systems (e.g., [319, 336]). To address these challenges, the responsible AI (RAI) community has called for more collaborative approaches to AI design and evaluation, where impacted stakeholders—with expertise in the specific real-world contexts that AI systems are deployed—play an active role in shaping the design and evaluation of AI systems [31, 97, 283]. However, much of this research has overlooked a critical opportunity to engage stakeholders in informing the underlying *measurement decisions* that undergird the design and evaluation of AI systems.

RAI research shows that AI systems often fail in deployment due to inappropriate *measurement assumptions*, made when translating from abstract, latent concepts (e.g., “fairness,” “toxicity”) to practical approaches and instruments for measuring those concepts—such as metrics, benchmarks, and LLM-based evaluators [201, 221, 254, 268, 346, 442]. For example, a widely used predictive tool in the US healthcare system was found to dramatically underestimate the health needs of Black patients, amplifying existing racial disparities in medicine [311]. This was a direct consequence of the decision to use patients’ incurred healthcare *costs* as a proxy variable for their *actual need* for care. Similarly, in social work, AI-based decision support tools (and their evaluations) are frequently built on invalid measurement assumptions from community and worker perspectives [220, 395]. For instance, while social workers aim to assess *immediate safety risks* to a child, AI tools are designed to predict the *long-term risk that a child will be removed* from their home by child services [138, 221]. Researchers studying NLP technologies have found that unjustified measurement assumptions in benchmark design often contribute to misleading evaluations [268], and those studying social media platforms have emphasized the need to re-examine the measurement assumptions underlying recommender systems (e.g., inferring users’ level of “interest” in certain video content based on their observed viewing time) [157, 402].

Measuring concepts such as “risk” [138, 221, 362], “fairness” of decisions [64, 149, 220, 387], “stereotyping” [34, 442], worker “well-being” [216, 259], or community members’ level of “need” for services like housing or healthcare [247, 311] necessarily requires making value-laden assumptions about how these theoretical concepts should be defined and operationalized into practical measurement instruments [5, 201]. Yet, in current practice, measurement assumptions are rarely interrogated or reasoned about with impacted stakeholders [221, 387]. Even when stakeholders are involved at early stages of AI design, discussions tend to stay at a broad conceptual level [247]. Non-technical stakeholders are commonly left out of the conversation when it comes time to decide how best to measure these concepts, in light of both project goals and practical constraints such as data availability [181, 222, 247, 404]. Without stakeholder involvement, the design of measurement instruments is often riddled with inaccurate assumptions that lead to downstream fairness, validity, and value alignment challenges [81, 153, 201, 202].

In this paper, we aim to guide nascent research on supporting more collaborative approaches to AI measurement by drawing lessons from the broader literature on stakeholder participation in measurement design. A growing body of research has argued that measurement design should not be a task left to AI practitioners alone [201, 221, 281, 387, 442]. While researchers in RAI have recently begun to explore methods to support more participatory approaches to AI measurement

(e.g., [64, 175, 274, 387]), these efforts currently remain fragmented, without a unifying lens to systematically understand challenges, opportunities, and differences across approaches. Importantly, the Responsible AI community is not alone in exploring these challenges. Beyond the context of AI, there are rich, existing traditions of stakeholder participation in the design of quantitative measurement approaches, across a range of disciplines—from the social and life sciences, to philosophy of science. As the RAI research community works to empower stakeholder participation in AI measurement, there is a critical opportunity (and responsibility) to draw lessons from and build upon this multi-disciplinary literature.

The remainder of this paper is organized as follows: Following an overview of existing research in AI measurement and collaborative measurement (Section 7.11), we provide a conceptualization of measurement as a *collaborative* design process (Section 7.3) with opportunities for stakeholder engagement across multiple tasks. We then conduct a multidisciplinary literature review of existing approaches to supporting stakeholder participation in each of these measurement design tasks (Section 7.12). Based on our review, we first overview the landscape of past methodological approaches (Section 7.5), and then synthesize a framework of methodological dimensions to characterize, distinguish, and assess design trade-offs (Section 7.6). Using our framework as a lens, we conduct an analysis of existing approaches to stakeholder participation in measurement, and highlight opportunities for future methodological innovation, to guide future RAI research (Section 7.7). We conclude with a discussion on how our framework can be adopted to support more collaborative approaches to AI measurement in practice (Section 7.14). Overall, by bridging lessons from a multidisciplinary literature, this paper provides a methodological framework and roadmap for future research and practice to advance more collaborative approaches to AI measurement.

7.2 Background

In this section, we first elaborate on the relationships between measurement and AI, briefly overviewing existing work that has called for adopting a measurement lens to better understand AI issues and failures (Subsection 7.2.1). We then provide context on historical motivations for empowering stakeholders in shaping measurement decisions, beyond the context of technology design, as well as emerging efforts within the RAI and adjacent communities to support stakeholder participation in measurement for AI applications (Subsection 7.2.2).

7.2.1 AI and Measurement

Developing, evaluating, and governing AI systems often requires translating abstract, unobservable concepts into *measurement instruments*—i.e., practical procedures and/or artifacts that operationalize a given concept into measure(s) [5, 201, 442]. Measurement instruments are pervasive across the AI development lifecycle, from the design of an AI system (e.g., choices of proxy variables aimed at measuring a latent concept in a predictive model [149]) to the evaluation of AI systems (e.g., metrics and benchmarks to measure model performance [442]) to tools intended to support AI governance (e.g., impact assessments that aim to measure potential downstream harms of AI systems [202]). Despite their pervasiveness, current practice in AI rarely acknowledges—much less, grapples with—the contestable, value-laden nature of

the measurement decisions being made. For instance, ML researchers and practitioners seldom explicitly define the underlying meanings of latent concepts that a measurement instrument is intended to operationalize, obscuring what exactly is being measured [442].

There is growing recognition in the RAI research community that inappropriate *measurement* assumptions during AI development often lead to real-world harms and failures to deliver promised benefits [81, 150, 201, 221, 281, 319, 346, 362]. In response, recent years have seen calls for increased attention to measurement and validity challenges in AI design (e.g., [81, 201, 336]), evaluation (e.g., [72, 149, 268, 442, 473]), and governance (e.g., [202]). For instance, researchers have described how *fairness* issues in AI systems can be better understood through adopting the language of measurement modeling [202]. Others have problematized how existing approaches to measuring “risk” in risk assessment algorithms used across high-stakes decision-making contexts (e.g., in child welfare, criminal justice) portray a misleading and biased understanding of the factors that shape stakeholder outcomes and decisions [362]. More recently, researchers have called for interrogating and improving the validity of generative AI evaluation practices through a framework grounded in measurement theory from the social sciences [442]. In this paper, we extend this body of scholarship by conceptualizing measurement in AI as a collaborative *design process*, and exploring how stakeholder participation throughout this design process can help overcome current measurement challenges.

7.2.2 Stakeholder Participation and Measurement Design

The measurement issues faced by the AI community closely parallel challenges other academic disciplines have long grappled with. In *Seeing Like a State*, James C. Scott revisits the formation of central governments to critically examine large-scale, state-driven social engineering projects [368]. These projects intended to make society more “legible” and governable by standardizing local practices into quantifiable measures. Scott argues that these standardization efforts reduced and distorted the complexity of local practices, making obsolete the practical and experiential knowledge (referred to as *metis*) that was central to the functioning of local communities. Stepping through several real-world examples of state failures—from the introduction of standardized forestry systems that led to long-term ecological degradation and soil depletion, to the implementation of new indicators for agricultural productivity that led to decreased farming efficiencies and food shortages—Scott concludes there is “no single, all-purpose, correct answer to a question implying measurement unless we specify the relevant local concerns that give rise to the question” [368].

Scott’s historical accounts of failures point to the importance of experiential, local knowledge and decentralized governance processes in developing interventions that align with the actual needs and practices of communities. These lessons are carried through today by pockets of research and practice, spanning a range of disciplines, that have explored techniques for stakeholder participation in measurement design (e.g., [8, 75, 121, 159, 278, 309, 381]). While measurement issues have been discussed using different terminology across these research communities, supported by different processes and techniques, they converge in their premise that measurement should be a design process involving collaboration between stakeholders with diverse expertise, beyond just technical experts.

In the RAI community, efforts to support stakeholder participation in AI measurement are still nascent. Past work in the area of “participatory AI” has rarely engaged stakeholders in interrogating and shaping AI measurement assumptions.

For example, while past work has engaged stakeholders in considering what model features should be used to *predict* an observable outcome (e.g., “cost of care”), stakeholders are not typically invited to question the measurement assumptions behind a predictive model—such as how well the outcome variable *measures* what it is intended to measure (e.g., “healthcare need”). However, a small but growing body of research has begun to explore methods for stakeholder participation specifically in AI measurement (e.g., [64, 175, 245, 274, 387]), such as workshop techniques that engage stakeholders in identifying metrics that can better capture their own experiences of fairness in recommender system evaluations [387] or interfaces aimed at soliciting stakeholders’ notions of statistical fairness in specific real-world contexts (e.g., [64, 274]).

Strengthening the responsible AI research community’s efforts to advance participatory approaches to AI measurement requires a systematic understanding of the challenges, opportunities, and differences across approaches. As the research community continues to explore this space, it is critical that we learn from and build upon the rich body of existing literature on supporting stakeholder participation in measurement design, beyond the specific context of AI development. Our work contributes to the literature on participatory AI by drawing lessons and insights from past work across disciplines, with the aim of guiding future methodological explorations in the RAI community.

As discussed in the next section, in this paper we conceptualize AI measurement as a *collaborative design process*. Inspired by traditions in co-design, participatory design, participatory research, and related areas, work in the broader area of participatory AI has called for a shift toward more collaborative approaches to stakeholder participation in AI development [33, 98, 411] that position non-technical stakeholders as genuine collaborators who bring complementary expertise to a design process and who directly participate in design decisions [112, 438]. Such an approach stands in contrast to more consultative approaches, where stakeholders may be consulted but not engaged as active designers, and power over design decisions is concentrated in the hands of AI developers [14, 98]. Our work addresses these broader calls with a specific focus on stakeholder participation in AI measurement design.

7.3 Measurement as Collaborative Design

In this section, we introduce the *Measurement Design Lifecycle* (Figure 7.1), a conceptualization of measurement in AI as a collaborative design process. Drawing on existing theoretical conceptualizations of both design [54, 98, 112, 365] and measurement [5, 8, 201], we conceptualize measurement design as a lifecycle including four main tasks, which are completed in an *iterative* (i.e., involving a cyclical process of ideation, feedback, and refinement) and *non-linear* (i.e., revisiting prior tasks and decisions, as needed) manner. As shown in Figure 7.1, the four tasks include: Goal and concept identification, Systematization, Operationalization, and Implementation. Our subsequent analyses of existing approaches to stakeholder participation (Section 7.5), along with challenges and opportunities (Section 7.7), are grounded in this conceptualization of measurement design. In line with past work on co-design [356], in the remainder of this section we refer to anyone engaged in the measurement design process as an active designer—including both AI and non-AI expert stakeholders—using the shorthand “designers”.

The design process begins with **Goal and Concept Identification**. This task involves determining the overall goals for measurement [98, 112], including how the

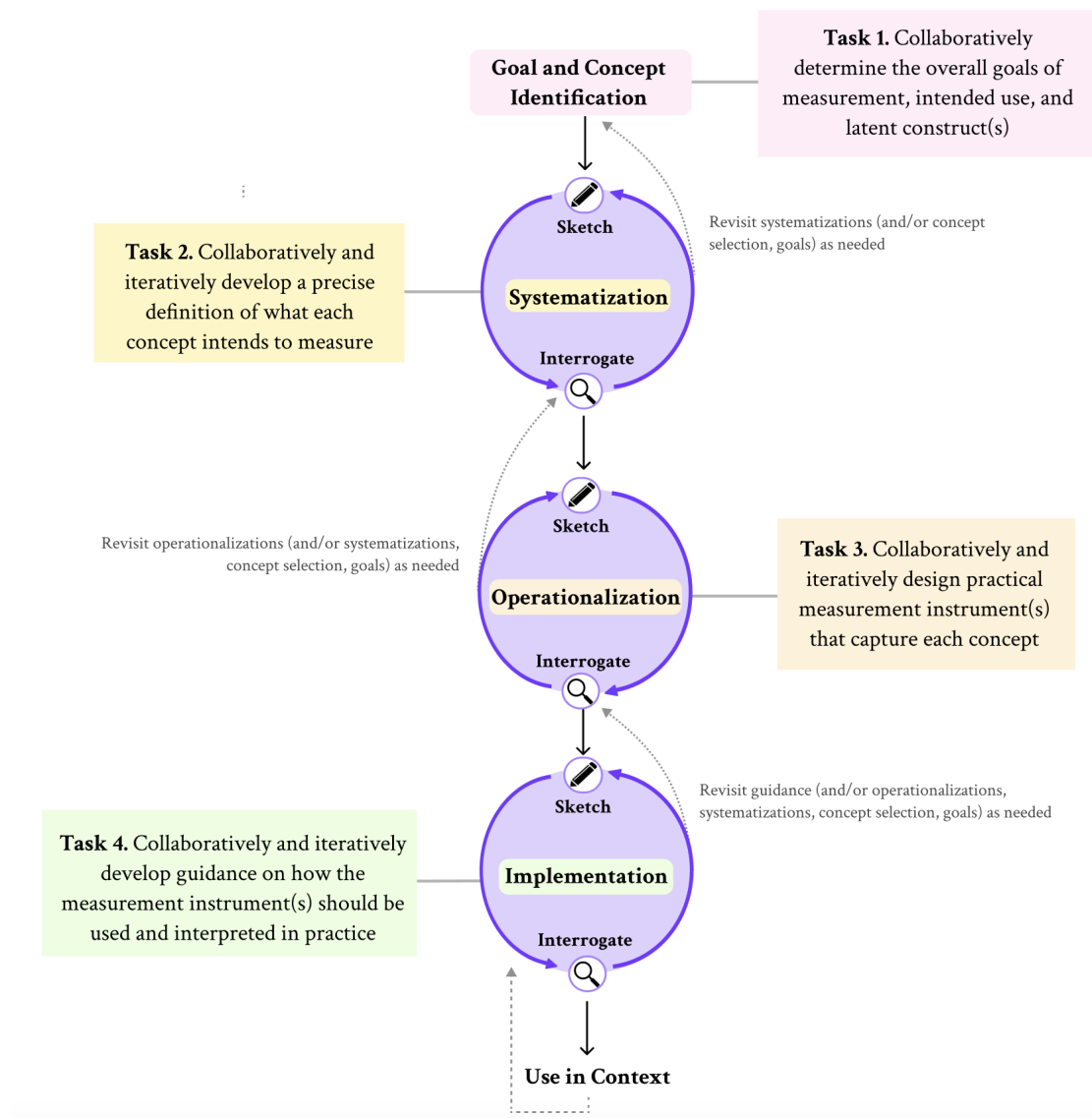


FIGURE 7.1: The *Measurement Design Lifecycle*, derived from a synthesis of existing frameworks of both design and measurement.

resulting measurements will be used. This task also involves identifying the latent concept(s) to be measured (e.g., “toxic language”, “wellbeing”, “equitable health-care service”). The next task, **Systematization**, involves taking these abstract, high-level concept(s) and more precisely defining what they are intended to capture [5, 442]. This process may involve any number of activities such as crafting abstract definitions, curating concrete examples to help define the boundaries of a concept, or constructing diagrams to clarify the interrelationships between subconcepts or related concepts. Importantly, as illustrated in Figure 7.1, systematization involves an iterative design cycle in which designers *sketch* out one or more potential systematizations of a concept, *interrogate* the systematization(s) (e.g., by attempting to poke holes in definitions and identifying opportunities for refinement), and then *iterate* on the systematization(s) as needed [54, 365]. During this process of attempting to define identified concepts, it is not unusual for designers to realize that they need to reconsider not only their systematizations, but also decisions made at the previous task, including what concept(s) to target or even what the overall goals of measurement should be [5]. For example, through systematizing a concept, a designer may realize that the high-level concepts they had identified initially are unlikely to help them achieve their overall goals, or that their goals may need reformulation. Figure 7.1 shows this potential for non-linearity via dotted, backwards-facing arrows between tasks.

Building on the systematizations developed in the previous task, **Operationalization** explores the design of practical measurement approaches and instruments to capture these concepts [5, 201, 319]. In line with theoretical notions of design as a “conversation” or “negotiation” with the materials at hand [276, 365], designers must ideate possible ways to operationalize a systematized concept in light of practical resource constraints (e.g., currently available or feasibly collectible data). As in the previous task, operationalization involves an iterative design cycle in which designers *sketch* out one or more ideas for potential operationalizations, *interrogate* these candidate operationalizations (e.g., by identifying potential validity concerns), and then *iterate* as needed. Once again, through the process of exploring potential operationalizations of a concept, designers may realize that they need to revisit decisions made at earlier tasks of the measurement design lifecycle [5]. For example, in the process of exploring potential proxy variables for their concept of interest, a designer might come to the conclusion that all possible candidate proxies raise unacceptable validity issues. In this scenario, the designer would ideally consider whether their original measurement goals need to be revised, or whether targeting alternative concepts might enable them to achieve their initial goals.

Finally, during **Implementation**, designers develop guidance for how the resulting measurement approaches and instruments should be *used in practice*, such as how measurements should be interpreted and used to guide decision-making [159]. Such guidance may be iteratively refined based on observations as a measurement approach is piloted or deployed in real-world contexts. As in previous tasks, this process of thinking through policies and guidance for use may prompt a designer to revisit design decisions made during prior tasks.

As argued above, addressing AI’s measurement challenges requires a shift toward treating measurement itself as an inclusive and collaborative part of the AI lifecycle, rather than a set of top-down decisions made exclusively by technical experts. **Each activity within this iterative design process presents an opportunity for stakeholder collaboration.** In the following sections, we draw insights from past efforts to engage non-technical stakeholders in AI measurement, and identify opportunities to advance research and practice in this space.

7.4 Methods

Making this turn toward more collaborative approaches to measurement in AI presents a critical opportunity for the Responsible AI community to *learn from* existing approaches across disciplines, both within and beyond the context of technology development. In this section, we describe how we developed our multidisciplinary corpus of papers, as well as how we used the measurement design lifecycle (Figure 7.1) to organize and analyze methodological approaches across disciplines.

7.4.1 Corpus Development

To develop our corpus of papers, we collected a set of research papers and grey literature (e.g., white papers from nonprofits) that describe or motivate methods for stakeholder participation in measurement. We searched for papers that support stakeholder participation for *any latent concept* and *any application*, not limited to technology design. Following prior reviews of participatory approaches to AI (e.g., [98]), we searched for papers that include *any form* of stakeholder participation. We consider stakeholders to include anyone who is asked to participate given specific expertise they bring to a measurement design process, beyond only quantitative or technical expertise, that the general population may not have (e.g., individuals with situated expertise relevant to the measurement application context). Appendix C.1 includes details on our search queries. To identify these papers, we first searched for relevant articles in conference and journal proceedings of the ACM Digital Library, IEEE Xplore Digital Library, Wiley Online Library, and SpringerLink. We selected these digital libraries to ensure the inclusion of major disciplines that have progressed work on measurement and participation, including the social sciences, statistics, computer science, HCI, health, education, and the physical sciences. We also used the same keywords (Appendix C.1) on Google Search to find relevant non-academic articles. We identified additional papers based on our team's knowledge of potentially relevant papers, and via forward and backward citation chaining (e.g., following citations within identified papers).

Our search yielded an initial corpus of 133 articles. We reviewed the query results to identify a final set of relevant articles that met our inclusion criteria, which was developed through iterative discussions around a sample of 20 papers from our initial corpus. First, to qualify for inclusion in our final corpus, papers needed to (1) engage stakeholders with the explicit end goal of *designing measurement approaches or instruments*, including any activity in the *Measurement Design Lifecycle* (Figure 1). For example, papers soliciting stakeholder notions of an unobservable construct without stating an intention to use this information to inform the design of measurements for that construct would be excluded. Similarly, a paper that supports stakeholder participation in identifying features that can help in the *prediction* of an observable variable (e.g., click-through rate, video watch time, number of words, occurrence of wildfires) with no stated intention to measure a latent construct, would be excluded. As a second inclusion criterion, papers needed to (2) propose or explore a *new technique* for stakeholder participation, or a new variation on an existing technique. Our third criterion was that papers must (3) engage people in shaping measurement approaches and instruments that are intended for common use by multiple individuals, rather than just one. This means that, for example, papers that engage people in personalizing an algorithm's behavior to their individual preferences (e.g., customizing a music recommendation algorithm or an algorithmic news feed) would not be included, given that the goal is not to design a measurement instrument that

can be used by others. Finally, our last criteria was that the paper must (4) include an empirical case study demonstrating the instantiation of the technique. The paper could include a singular end-to-end case study or multiple case studies that demonstrate different portions of the proposed technique. Papers that only present a high-level conceptual overview of the technique without an empirical demonstration of how the technique was used in practice were removed. Appendix C Figure C.1 details the number of papers removed from the corpus for each exclusion criterion.

Our final corpus included 39 articles published across a range of disciplines, including HCI, design, ML, political science, statistics, psychology, healthcare, education, dentistry, agriculture, sustainability, ecology, food safety, and community planning. Table 7.1 includes a distribution of years and higher-level disciplines covered in our final corpus, and Appendix Table C.3 includes the 39 articles.

Item	Count of Papers
Year	2000 (1), 2001 (1), 2003 (1), 2006 (1), 2008 (2), 2009 (1), 2010 (2), 2011 (1), 2014 (2), 2015 (1), 2016 (1), 2017 (1), 2019 (1), 2020 (4), 2021 (4), 2022 (3), 2023 (4), 2024 (5), 2025 (2)
Discipline	Health and Medicine (11), Social Sciences and Policy (11), HCI and Design (8), Sustainability and Agriculture (4), Responsible AI (2), Machine Learning (2), Philosophy of Science (1)

TABLE 7.1: Metadata about the papers in our corpus.

7.4.2 Analysis

We conducted two phases of analysis. First, we analyzed the 39 papers to understand the breadth of approaches represented in our dataset. This informed our overview of existing approaches (Section 7.5). Drawing on this analysis, we synthesized a framework of dimensions and use it as a lens to enrich our analysis of the 39 papers in our corpus (Section 7.6 and 7.7). Below, we elaborate on our analysis approach.

Initial analysis of approaches. Our first goal in analyzing the 39 papers was to understand the breadth of methodological approaches proposed across disciplines, challenges, and trade-offs across approaches, and opportunities for methodological innovation. At least two members of our researcher team reviewed each paper in the corpus. One researcher took multiple passes of each paper to characterize the overall approach(es) the paper used; the researchers’ goals for using that approach; and the measurement design task(s) the approach(es) support. In addition, for each paper, the researcher identified strengths, challenges, and trade-offs of the approach(es), with respect to its potential to support truly *collaborative* approaches to measurement design, beyond stakeholder *consultation* [14, 98]. The second researcher then reviewed each paper to identify additional insights and points of disagreement with the first researcher. The two researchers discussed and resolved any points of disagreements, revising the analysis accordingly.

Synthesis & application of a new framework Our analysis of existing approaches informed the development of a set of eight dimensions that collectively characterize a *methodological design space* for participation in measurement (Section 7.6). In particular, using a constant comparative approach [151], two researchers iteratively discussed definitions and examples of the eight dimensions, drawing on the review of existing approaches and on frameworks for participation in design more broadly (e.g., [98, 112, 345]). Using these eight dimensions, we systematically analyzed methodological design decisions across approaches in our paper corpus (Section 7.7). First, one researcher applied the framework to code whether each approach in the corpus rated “low,” “medium” or “high” on each of the applicable eight dimensions. We include examples of the types of activities scoring “lower” or “higher” on a given dimension in Section 7.6. The second researcher then reviewed the coding to identify any areas of disagreement; disagreements were resolved through discussion with the first researcher.

7.5 Approaches to Participation Across the Measurement Design Lifecycle

To concretely illustrate how stakeholders can be engaged at each phase of measurement design, we begin by presenting a small sample of approaches from our corpus. Our goal in this section is to sensitize the reader to a diversity of concrete approaches proposed by prior literature, before abstracting to higher-level dimensions that characterize these approaches in the next section. We include the count of papers in our corpus that supports stakeholder participation in a given measurement task. Of the 39 papers in our corpus, only eight papers support more than one task.

Task 1: Collaborative Identification of Measurement Goals and Concepts (4 papers). Prior approaches have engaged stakeholders in shaping overall measurement goals through surveys, interviews, and both individual and group co-design activities. For instance, the *Community KPIs* approach (P7) involves local citizens in defining measurement goals and corresponding indicators (KPIs) to help evaluate smart city IoT deployments [159]. In the first stage of this process, citizens participate in a workshop where they collectively define what “success” for a given deployment looks like from their perspectives. As preparation for this brainstorming activity, workshop participants begin by discussing potential benefits and harms of the IoT use case, including undesirable outcomes and potential to positively impact citizen lives. In subsequent phases of measurement design, practical approaches are designed to monitor the success criteria that emerge from this phase.

Task 2: Collaborative Systematization of Concepts (17 papers). Prior work has explored a range of approaches to collaboratively define high-level concepts. For example, some approaches engage stakeholder expertise to understand *individuals’ experiences and boundaries* around a concept, while others focus on *situating the concept within broader systems*:

- **Iteratively co-developing a grounded theory:** Paper P31 engaged multiple stakeholder groups—individuals with lived expertise in poverty, practitioners from a UK-based anti-poverty charity, and academic scholars with expertise in wellbeing—in co-producing a theory for “thriving” as a foundation for the design of measures of financial hardship in the UK [8]. To collaboratively develop this theory, individuals across the three stakeholder groups first conducted relational interviews with one another to better understand what

“thriving” meant to someone outside of their own stakeholder group. The researchers then analyzed the interview transcripts and presented their initial systematization of “thriving” back to stakeholders. Stakeholders engaged in debate and collaborative refinement to ensure the systematization accurately reflected their notions of “thriving.”

- **Collaboratively constructing system diagrams:** In paper P33, researchers engaged stakeholders in systematizing the concept of “wellbeing” to guide future efforts to develop a national wellbeing index for Ireland [166]. The researchers conducted multiple workshops with stakeholders—including older adults, younger adults, children, and attendees of a wellbeing conference. Given the researchers’ goal of creating an index that would be useful in *guiding action* to improve national wellbeing, they adopted a systems science approach (Interactive Management) to understand systemic influences on wellbeing. For example, in one activity, stakeholders collectively created a map of *systemic barriers* to wellbeing that they perceived, along with causal interrelationships among these barriers. This system understanding guided later efforts to operationalize an actionable index.

Task 3: Collaborative Operationalization of Concepts (28 papers). Prior work has explored a range of approaches to engage stakeholders in operationalizing unobservable concepts into measurement instruments. Beyond the sorts of structured elicitation approaches used in past HCI and RAI research—such as pairwise comparison tasks to elicit stakeholders’ fairness notions [64, 392]—past work has explored more collaborative and flexible modes of stakeholder participation:

- **Supporting open ideation and discussion of potential operationalizations:** The *Everyday Peace Indicators* approach (P22) engages local community members in generating ideas for observable indicators that can be used to measure and monitor “peace” in their specific community contexts [278]. Community members engage in focus groups, where they are asked to generate a large list of indicators that they *already* use to informally gauge the current level of “peacefulness” in their communities—such as the frequency of gunshots overheard, the level of responsiveness of security/police to reports, or increases in unlicensed drinking establishments. In a next round, community members then collaboratively refine the generated list of indicators, adding and removing items as they see fit based on discussion, and voting on the importance of indicators.
- **Helping stakeholders interrogate the quality of potential operationalizations:** Some approaches have explored how to support stakeholders in more effectively interrogating the quality of potential operationalizations, to inform iteration and refinement. For example, prior work has explored the design of interactive interfaces and “what if” simulations (P21, P35, P36) to help stakeholders understand the potential implications of specific operationalization decisions for validity [175], fairness and value-alignment [64, 274], or usefulness [175].

Task 4: Collaborative Implementation of Measurement Instruments (3 papers). In a later step of the *Community KPIs* approach mentioned above [159], local citizens participate in a workshop to design an “observational system” to support the actual *collection of measurements*, using the measurement instruments they previously co-designed (P7). The workshop involves community representatives, optionally

alongside researchers and designers, in deciding how co-designed measurement approaches should be implemented in practice (e.g., at what intervals measurements should be collected and who will be accountable for collecting measurements).

7.6 Dimensions of Collaboration in Measurement Design

In this section, we present a framework of eight dimensions to characterize, assess, and compare across different approaches to supporting stakeholder collaboration in measurement. Each dimension captures a design property that can either facilitate or hinder a truly *collaborative* approach to measurement design. The first six dimensions can be used to characterize participation in specific measurement design tasks, such as systematization or operationalization. They include the following:

- **Expressiveness:** How much freedom are stakeholders provided to share their expertise, in a given measurement design task? For example, are participants able to express their perspectives and ideas in any form they desire (higher expressiveness), or are they asked to “fit” their responses into pre-defined categories and options (lower expressiveness)?
- **Directness:** How much interpretation must technical experts do on stakeholders’ behalf to make their inputs usable? For example, can technical experts simply take the final outputs of participatory activities into the next design task as-is (higher directness)? Or must they make additional assumptions about how to interpret or implement stakeholder inputs after stakeholders have left the room (lower directness)?
- **Evaluability:** To what extent are stakeholders empowered to evaluate the quality of their own or others’ ideas? For example, are they presented with information and representations that enable them to assess the alignment of an operationalization against their own unique expertise and knowledge of a domain (higher evaluability)? Or are they only presented with decontextualized, aggregate statistics that are difficult to connect with their own experiences (lower evaluability)?
- **Iterativeness:** To what extent are stakeholders able to iterate on their responses? For example, are stakeholders given an opportunity to assess the quality of a proposed operationalization, and then to iterate as needed (higher iterativeness)? Or are stakeholders asked to develop their ideas in “one shot” (lower iterativeness)?
- **Collectiveness:** To what extent do participants have opportunities to learn from other stakeholders’ expertise? For example, are stakeholders able to discuss, critique, and build upon others’ ideas (higher collectiveness), or do stakeholders contribute independently, in isolation from one another (lower collectiveness)?
- **Scalability:** How readily can a given activity be used to collect, interpret, and use larger numbers of participant inputs—whether inputs from larger numbers of people, larger numbers of inputs per person, or a combination of these factors? For example, can technical experts automatically collect and use responses from large numbers of stakeholders (higher scalability)? Or are there human constraints, such as a requirement for human facilitation or extensive

deliberation among stakeholders, that work against large-scale participation (lower scalability)?

While our first six dimensions capture properties of specific measurement design tasks, the last two dimensions characterize the overall approach to participation taken in a paper, which may span multiple design tasks:

- **Comprehensiveness:** To what extent are non-technical stakeholders involved throughout the measurement design lifecycle? Are stakeholders involved across every design task (higher comprehensiveness), or only in certain tasks, while decisions in other design tasks are made on their behalf (lower comprehensiveness)?
- **Non-linearity:** To what extent can stakeholders revisit prior decisions, made in earlier tasks within the measurement design lifecycle? For example, during operationalization, is there space to reconsider the overall goals for measurement or the definitions of concepts (less linear), or are participants required to proceed with operationalization of the previously identified concepts (more linear)?

7.7 Methodological Observations & Opportunities

In this section, we demonstrate how the framework described in Section 7.6 can be applied to systematically characterize and compare different approaches to participatory measurement design. In doing so, we identify a set of eight observations about existing methodological approaches in our corpus. Drawing implications from these observations, we present opportunities to develop more *collaborative* approaches to stakeholder participation in measurement, including by (1) supporting more **comprehensive** and **non-linear** participation across the measurement design lifecycle (Subsection 7.7.1); (2) engaging stakeholders in more **iterative** design processes, supported by mechanisms that support greater **evaluability** of measurement design ideas and **collectiveness** in developing and evaluating ideas (Subsection 7.7.2); and (3) exploring ways to avoid sacrificing **expressiveness** of stakeholder participation in favor of supporting **scalability** or **directness** (Subsection 7.7.3). Across this section, we include counts (X/Y, Z%) to indicate the proportions and percentages of papers in our corpus represented by each finding.

7.7.1 Opportunities to Support Participation Across the Measurement Design Lifecycle.

We found that most papers in our corpus engaged stakeholders in isolated design tasks such as operationalization, without opportunities to revisit past design decisions and assumptions. Furthermore, very few papers engaged stakeholders in “big picture” decisions such as determining the overall goals of measurement in a given context and shaping how a measurement instrument will ultimately be used in practice. Based on these observations, we discuss opportunities for future research to support more **comprehensive** and **non-linear** modes of participation in measurement design.

OBSERVATION 1: Participation in isolated measurement design tasks, with no opportunity to revisit past decisions and assumptions.

In terms of **comprehensiveness**, the vast majority of papers support stakeholder participation in only one of the four tasks of measurement design (31/39, 79%). There are four exceptions (P7, P19, P36, P37) which take a more comprehensive approach, all of which support three or more tasks. Of these four papers, only one (P36) supports a **non-linear** approach, supporting stakeholders in iterating between tasks. This aligns with a broader trend among papers in our corpus: All but one of the papers in our corpus followed a *linear* design process, where stakeholders had no opportunities to revisit and refine decisions made in prior tasks. This includes revisiting both their own prior decisions and those made by researchers, such as what to measure or which measurement goals and concepts to use. An exception is P36, in which stakeholders were engaged in revisiting their prior systematizations (Task 2) based on observations at the operationalization stage, as well as revisiting their prior operationalizations (Task 3) based on observations from the implementation of those measurement instruments (Task 4).

OBSERVATION 2: Minimal coverage in the earliest and latest tasks

We also observed that the majority of papers in our corpus engage stakeholders in collaborative operationalization (28/39, 72%), followed by collaborative systematization (17/39, 44%). In contrast, relatively few papers support the collaborative identification of measurement goals and concepts (4/39, 10%) or collaborative implementation (3/39, 8%). Only four papers (P2, P7, P36, P37) involved stakeholders in deciding *the goals of measurement*. Similarly, only three papers (P7, P19, P36) involved stakeholders in later-stage decisions around *how* to apply measurements to a use context (e.g., what thresholds to consider, how to interpret measurement outcomes, how often to collect measures).

OBSERVATION 3: Approaches tend to engage stakeholders in either systematization or operationalization, but not both.

Out of the 17 papers that involved stakeholders in systematization, most did not involve stakeholders in *operationalizing* the systematized concepts afterwards (10/17, 59%). Instead, technical experts often took on the task of translating these concepts into implementable measurement approaches, without further stakeholder involvement. In the six papers (P11, P17, P19, P31, P37, P39) where stakeholders are involved in both systematization and operationalization, they are often only involved in ideating operationalizations without opportunity for feedback and refinement (P11, P17, P19, P31) or only involved in validating operationalizations that technical experts made themselves without opportunity for ideation of new operationalizations (P37, P39). One paper (P39) presents an exception, where stakeholders are involved in the full process of ideation, assessment, and iteration in both the systematization and operationalization tasks. Meanwhile, out of the 28 papers that involved stakeholders in operationalizing a concept, most immediately moved ahead to operationalization rather than engaging stakeholders in a dedicated systematization step to define and clarify the concepts to be operationalized (21/28, 75%).

IMPLICATIONS & OPPORTUNITIES

Overall, these findings suggest opportunities for future methods to support more **comprehensive** and **non-linear** stakeholder participation across the measurement lifecycle. Supporting flexible participation *across* design tasks has potential to yield fundamentally different design outcomes. When technical experts design measurement approaches, the process is inherently messy and non-linear. Insights from a given measurement design task often inform revisions to prior design decisions [5, 150]. Future work should enable stakeholders to gather and act upon similar insights. For example, in the process of trying to operationalize a systematized concept, stakeholders may realize that there are important gaps in their systematization, leading them to reconsider what it is that they truly wish to measure. Similarly, through the process of exploring potential operationalizations, stakeholders may realize that it is not practically feasible to develop an operationalization that they would consider adequate to capture their notions of a given concept [150]. In this scenario, rather than simply pushing ahead with potentially problematic operationalizations, they may wish to “pivot” by revisiting their overall goals for measurement, which specific concepts to measure given these goals, or how exactly to define those concepts. It remains an open question how best to support and encourage fluid transitions between different measurement design tasks, while minimizing the time and effort costs of non-linear design exploration (cf. [409]).

Relatedly, as *Observation 2* describes, papers in our corpus rarely engaged stakeholders in more fundamental decisions about *whether and why* to measure in the first place and *how* measurement approaches should actually be used in practice. These observations mirror trends documented by Delgado et al. in the broader literature on “participatory AI” [98]. As our community advances new methods for stakeholder participation in measurement design, it will be critical to ensure that these larger decisions are on the table for discussion.

Finally, these findings speak to opportunities to develop methods that more closely integrate stakeholder participation across systematization and operationalization. Past Responsible AI research has observed that ML researchers and practitioners often entirely skip systematization as a step in measurement. Researchers have called for greater attention to systematization, and have suggested that systematization represents a critical opportunity for stakeholder participation [442]. We observe a similar trend among papers in our corpus that engage stakeholders in operationalization (*Observation 3*): Papers that engage stakeholders in operationalizing a given construct often do not engage them in first systematizing the concept. We also observe that, among papers that do engage stakeholders in systematization, this is often done *in place of* engaging stakeholders in operationalization. Yet past research suggests that designing more valid measurement instruments that can effectively capture stakeholders’ notions of a given concept requires continued stakeholder involvement in operationalization—a stage at which many consequential design decisions must be made, for example, to account for data limitations and other practical constraints [150, 175, 221, 311, 346]. Even with a comprehensive understanding of non-technical stakeholders’ beliefs and preferences regarding an unobservable concept, operationalizing this understanding into practical measurement instruments presents many opportunities to misrepresent stakeholder perspectives. Future work should explore ways to support stakeholder participation in *both* systematization and operationalization as part of an integrated, iterative process.

7.7.2 Opportunities to Support Iterative Ideation & Assessment in Measurement Tasks.

Overall, we observed that papers in our corpus often do not support stakeholders in *iterating* within a given design task. For instance, while several papers in our corpus provided space for stakeholders to weigh measurement design alternatives, few empowered them to meaningfully assess the quality or feasibility of design ideas. Even fewer provided opportunities for stakeholders to use such assessments to iteratively update and refine their ideas. Based on these observations, we discuss opportunities for future research to support more **iterative** modes of participation.

OBSERVATION 4: Stakeholders are rarely empowered with knowledge of relevant practical constraints to generate design ideas that are implementable in practice

Empowering stakeholders to make design decisions that can directly be used in practice (**directness**) requires that they are sensitized to relevant data considerations and other practical constraints when ideating and assessing design ideas. Among papers in our corpus, we found that only one paper (P36) sensitized stakeholders to relevant design constraints during systematization and operationalization. We observed some cases where stakeholders raised potential practical constraints through the process of ideating potential operationalizations, for example, the feasibility of collecting specific types of data. However, stakeholders were not provided with an opportunity to iteratively refine the operationalized concepts based on these new considerations that they raised through ideation. For example, in P16, stakeholders engaging in a workshop discussion of potential operationalizations of trout “welfare” raised critical considerations around the *feasibility* of specific measurement ideas, given practical implementation constraints like the feasibility of measuring water quality as an indicator in every fish farm as an indicator of trout welfare. While these discussions created an opportunity for stakeholders to grapple more deeply with real-world constraints around measurement, the lack of explicit feedback from researchers on what these data constraints include or how to go about addressing them left the workshop without actionable outcomes (though the workshop advanced on other fronts). In other cases, technical stakeholders reflected on the need for future work to better sensitize stakeholders to data constraints, after finding that stakeholders’ ideas for how to operationalize a given concept cannot be directly used in practice. For example, P19 involved stakeholders in a workshop to identify definitions, goals, and measurements for “fairness.” To help assess these measurement ideas, participants were asked to discuss “tradeoffs and challenges” together. This led participants to discover two competing desires for fairness measurements, but lacked guidance to understand how this trade-off should be accounted for when designing a measure. Similarly, in P04, researchers invited non-technical-stakeholders (e.g., teachers, students) to workshops to design evaluation metrics for datasets used in AI deployments in education. Upon reflecting on their workshop study, the researchers discussed how non-technical stakeholders were unable to articulate concrete ideas for new metrics, despite recognizing limitations in existing status quo metrics used in their context, and technical stakeholders were unable to make teachers’ measurement design ideas into actionable data specifications.

OBSERVATION 5: Stakeholders receive limited support in evaluating design ideas

To support design iteration, it is critical that stakeholders are empowered to evaluate the quality of their own or others' design ideas (**evaluability**). However, in the context of designing measurement instruments, it can be challenging for stakeholders to fully assess design ideas without additional support. For example, when comparing a set of alternative operationalizations—whether quantitative metric definitions or natural language prompts for LLM-based metrics—it can be challenging for stakeholders to anticipate the consequences of different operationalization decisions in downstream application contexts. Many papers in our corpus (25/39, 64%) provided space for stakeholders to weigh measurement design alternatives. Yet we found that relatively few papers (3/39, 8%) empowered stakeholders to evaluate the quality of measurement design ideas based on their unique knowledge and expertise (**evaluability**).

A handful of papers in our corpus presented stakeholders with interactive representations aimed at helping them explore potential downstream implications of alternative operationalization decisions (3/39, 8%). P21 and P35 provided participants with interactive visualizations that allowed them to examine how alternative statistical fairness metrics would behave across demographic groups, using either real or simulated data. In P35, participants could explore how different operationalization decisions would impact fairness judgments both at a distributional level (e.g., examining rates of false positives and false negatives across demographic groups) and through instance-level comparisons (e.g., allowing stakeholders to examine to what extent individuals with “similar” demographic profiles are treated similarly). In both P21 and P35, these interfaces supported participants in understanding how well different operationalizations aligned with their own notions of what it means to be statistically fair in a given real-world decision task. Similarly, P36 used past log data from students' interactions with educational software to make the implications of operationalization decisions tangible to participants. In this work, math teachers iteratively co-designed a set of automated “detectors”, each intended to measure a specific aspect of student learning and engagement—for example, whether a given student is “making careless errors” or is “unlikely to learn a given skill unless a teacher intervenes to provide extra help”. To examine how different operationalizations behaved, math teachers viewed live replays of students completing math problems in the software, annotated with live outputs from each detector. This allowed teachers to directly compare the detectors' interpretations with their own expert interpretations based on their observations of students' work. Through this process, teachers discovered that potential operationalizations that had seemed reasonable in concept had major validity issues in practice. They collaborated with a technical expert to rapidly iterate and test alternative operationalizations.

Another approach papers took to help stakeholders evaluate design decisions was to facilitate perspective sharing among participants. In several of the papers in our corpus, stakeholders were given opportunities to hear other participants' thoughts on a systematization or operationalization (15/39, 38%). While such perspective sharing has the potential to sensitize individual participants to a wider range of measurement design considerations, in many cases, such perspective sharing was conducted in a constrained manner. For example, five papers used versions of the Delphi technique, where each participant was able to view other participants' aggregate rankings of the importance of different subconcepts or indicators (P2, P15, P25, P29, P30), and then to update their own rankings as they saw fit. Such approaches can help participants get a sense of others' priorities among

presented options, but do not support them in understanding the *underlying reasons* for differences in perspective. A smaller subset of papers (2/39, 5%) explored more **collective** and **expressive** approaches—involving back-and-forth discussions where participants are guided in critiquing and building upon each other’s ideas—in turn supporting **higher evaluability**. For example, in P34, stakeholders engaged in synchronous discussion in a workshop setting to collaboratively develop and refine a systematization of “wellbeing” for their specific context.

OBSERVATION 6: Stakeholders have limited opportunity to iterate on designs.

We also observed that, amongst the 24 papers where stakeholders received some degree of support in assessing measurement design ideas, only nine papers provide stakeholders with an opportunity to act upon their insights by **iterating** on prior design decisions (9/24, 38%). In cases where stakeholders did have an opportunity to iterate, they were constrained in how they could do so. For example, while several of the papers in our corpus supported stakeholders in updating their stated *prioritization* among ideas for operationalizations, participants were not given an opportunity to express ideas for new measures. Five papers in our corpus (P07, P11, P20, P34, P36) present exceptions, each demonstrating an approach to support stakeholders in iterating on systematizations or operationalizations while supporting high expressivity. For example, in P34, researchers held discussion groups where stakeholders were invited to debate and revise systematized concepts representing “wellbeing,” allowing for rapid iteration within a single facilitated group session. In P20, the researchers developed an online environment where Wikipedia editors could come together and asynchronously critique, debate, and revise each others’ decisions about how to operationalize the concepts of “vandalism” and “damaging edits,” when collaboratively developing AI evaluation datasets.

IMPLICATIONS & OPPORTUNITIES

Overall, these motivate the need for future work to explore methods that provide non-technical stakeholders with greater opportunities for revision and iteration (*Observation 6*), informed by knowledge of relevant design constraints (*Observation 4*) and support in evaluating the quality of measurement design ideas (*Observation 5*). To do this, future work should explore approaches to better bring out the expertise of both technical and non-technical stakeholders, to enable closer collaboration and mutual learning amongst stakeholders. For example, *Observation 4* describes how non-technical stakeholders are often not sensitized to practical design constraints (e.g., data limitations) when systematizing or operationalizing a given concept. Future methods should leverage technical experts’ knowledge of these practical constraints to *help guide* stakeholder ideation in more productive directions, so the ideas stakeholders generate have a greater chance of actually being implemented in ways that stakeholders would consider valid. For example, future research can draw inspiration from existing efforts to scaffold effective ideation of concepts for AI applications, where researchers develop resources to help stakeholders quickly understand relevant technical parameters and constraints, so that they can more effectively assess AI design decisions and ideate alternatives [247, 386, 463]. In addition, future research might draw inspiration from past research that scaffolds conversations between technical and non-technical stakeholders to identify practical data constraints that could impact the feasibility of an AI system design concept [219]. Similarly, *Observation 5* describes a small handful of papers that demonstrate ways to support

stakeholders in evaluating measurement design ideas by drawing upon their own unique knowledge and expertise around the concept to be measured. Future work should continue to build upon these approaches, for example, by making potential downstream consequences of specific measurement design decisions visible to stakeholders, or by scaffolding stakeholders in *collaboratively* evaluating measurement design ideas and learning from each other's perspectives.

7.7.3 Opportunity to Support Larger-scale Participation Without Sacrificing Expressiveness

Overall, we found that papers in our corpus tend to trade-off the expressiveness of participation (i.e., the degree of freedom stakeholders are provided to share their expertise) for the directness and scalability of an approach. We discuss opportunities for future research to avoid these apparent trade-offs through methods for stakeholder participation that achieve better balances of **expressiveness**, **directness**, and **scalability**.

OBSERVATION 7: Stakeholders are not often invited to validate technical experts' interpretations of their inputs to design.

Several papers in our corpus support highly *expressive* modes of stakeholder participation (29/39, 74%), in systematization (11/29, 38%), operationalization (10/29, 34%), or both (5/29, 17%). Such papers provide participants greater flexibility in how they express their ideas, rather than constraining their responses into pre-defined response formats. However, particularly during operationalization, the outputs of more expressive activities often required significant additional interpretation to translate them into readily usable operationalizations. Importantly, we observed that technical experts tended to perform this translation of stakeholders' inputs into usable operationalizations *on stakeholders' behalf*, without inviting them to validate the accuracy or quality of these interpretations. In other words, we observed that approaches tended to trade off **directness** for **expressiveness**. Of the 15 papers in our corpus that support highly expressive modes of participation in operationalization, 9 were coded as having low directness. Meanwhile, of the 14 papers that support relatively *direct* forms of participation in operationalization, 10 were coded as having low expressiveness.

OBSERVATION 8: More readily scalable approaches tend to have limited expressiveness.

Across tasks, we observe that approaches aimed at supporting larger-scale participation in measurement design tend to sacrifice expressiveness to achieve this goal—mirroring tensions between scalability and expressiveness that have been discussed in the context of participatory AI more broadly [345, 466]. Of the 11 papers that have highly scalable activities in our corpus, all sacrificed stakeholders' ability to freely express their expertise (lower expressiveness)—whether by constraining responses to pre-defined categories in the form of survey response options (P10, P15, P20, P25-P30) or options for a pairwise comparisons activity (P11, P12). Notably, while no papers support both highly expressive *and* highly scalable approaches, a few papers in our corpus support relatively high levels of expressiveness while also supporting greater scalability (P10, P15, P20, P21, P29, P35). For example, while P10 uses a large-scale online survey to solicit participants' notions of desirable LLM

behavior, the survey design combines closed-ended response options with regular opportunities for participants to freely express their thoughts through open-ended response options. As we will discuss below, advances in LLMs introduce opportunities to make effective use of such free-form responses at scale. As another example, P20 introduces a platform that supports large-scale online collaboration in designing community-specific AI evaluation datasets. However, rather than having participants independently contribute labeled data points, the platform allows them to openly communicate with each other to deliberate and coordinate the process. For instance, at any point in the process, participants have the option to collectively revisit what label definitions they want to use.

IMPLICATIONS & OPPORTUNITIES

Overall, our analyses suggest an opportunity for future work to explore methods that avoid sacrificing **expressiveness** in favor of **directness** and **scalability**. First, we believe that apparent trade-offs between expressiveness and directness can largely be avoided through alternative methodological decisions. For example, approaches that support stakeholders in iteratively validating and refining technical experts' ideas for operationalizations, can preserve expressiveness while reducing the risk of misinterpretation in the process of making the outputs of early design activities usable.

We also see potential for future approaches to support larger-scale participation without overly constraining stakeholders' ability to express their ideas and perspectives. For example, building on recent advances in large language models (LLMs), future work could explore automated approaches to flexibly elicit people's nuanced preferences and ideas via multi-turn, chat-based conversations with individuals (cf. [263]) or LLM-facilitated online group discussions [340, 384]). Unlike pairwise comparisons or multiple choice surveys, where stakeholders are forced to fit their preferences into pre-defined options [347], such approaches have potential to support more expressive stakeholder contributions via free-form responses and discussions. These approaches aim to scale up some of the roles that human facilitators play in the context of small-group activities such as workshops: gently steering conversations and adaptively asking participants follow-up questions. In the context of measurement design, automated facilitation could be designed to help stakeholders iteratively clarify their ideas and provide additional information needed for measurement design, thus supporting greater directness in addition to expressiveness.

Future approaches could also build upon recent work in HCI on LLM-supported workflows and interfaces that lower the time and effort needed to iteratively, rapidly prototype measurement instruments for AI evaluation (e.g., [99, 254, 255, 374]). While much of this work has focused on supporting technical stakeholders' prototyping efforts, future research could explore and test adaptations to support and scaffold effective participation of *non-technical* stakeholders with important expertise to contribute to measurement design [412, 413]. Overall, future research should continue to expand these lines of work by exploring where carefully designed (semi-)automated facilitation and scaffolding could support more expressive and direct forms of stakeholder participation in measurement design at scale.

7.8 Discussion

Growing research in HCI and responsible AI (RAI) has called for re-examining AI failures through a measurement lens—examining how downstream harms can stem from inappropriate assumptions around how complex concepts (like “fairness,” “well-being,” or “risk”) should be conceptualized and operationalized into quantitative measurement instruments [81, 150, 201, 222, 281, 319, 362]. Addressing these measurement issues, and designing AI systems that truly benefit stakeholders, requires that stakeholders themselves are involved in shaping measurement assumptions and design decisions. While the HCI and RAI communities have explored a range of approaches to supporting stakeholder participation in AI design—including in the design of AI interfaces, models, and policies—these approaches have seldom invited stakeholders to interrogate and help shape the measurement assumptions that underpin them, which often remain opaque to stakeholders.

A nascent body of research from the HCI and RAI communities has begun to explore how stakeholders can be effectively engaged in AI measurement design. However, such efforts remain relatively fragmented. Our paper introduces language, concepts, and a methodological framework to guide the research community in more systematically exploring this new problem space—drawing lessons from decades of existing work on supporting stakeholder participation in measurement design, both within and beyond the context of AI. We apply our framework to analyze methodological trends across these existing approaches. Our analysis reveals a range of opportunities to advance approaches to stakeholder participation. By examining emerging work on participation in AI measurement design through the lens of our framework, we find that current approaches tend to constrain stakeholders to a linear process, and do not support stakeholders in iteratively exploring, evaluating, and refining design ideas. In addition, approaches in AI often restrict stakeholder contributions to isolated decisions within the broader measurement design lifecycle, and constrain what ideas and perspectives they are able to express (e.g., restricting participation to choices among a small set of pre-determined options).

Our analysis suggests several opportunities for future methodological innovation to overcome these limitations, while addressing the unique challenges of designing for AI. We believe that the CHI community, with deep expertise in innovating on design approaches and tools, is particularly well-suited to drive methodological advances in this space. Looking forward, we envision that our framework can help guide future researchers and practitioners interested in designing or assessing among alternative approaches to participatory measurement design. In practice, applying the framework and developing a strategy for stakeholder participation surfaces additional decision points, in light of practical considerations and real-world constraints in a given application context. We discuss some of these below, bringing discussions from prior literature on participation and participatory AI in conversation with our framework.

When involving stakeholders in measurement design, it is important to consider what complementary forms of expertise different stakeholder groups can bring to the process. Researchers and practitioners should first reflect on gaps in their own knowledge and identify stakeholder groups with expertise relevant to the design process, drawing on existing guidance from participatory research and design (e.g., [318, 438]). For example, when designing or evaluating AI applications for healthcare, the design of valid measurement approaches can require not only the *technical expertise* of AI practitioners, but also the *lived expertise* of impacted community members, the *frontline expertise* of the healthcare professionals who interact

with them day-to-day, and the *scholarly expertise* of researchers who study systemic biases in healthcare. Furthermore, initial discussions with stakeholders can reveal additional relevant stakeholder groups that a team may have initially overlooked.

Participatory approaches to measurement—like participatory design and research approaches more broadly—are not a monolithic practice, but rather, an orientation towards centering stakeholder expertise and ways to elevate it through the process of design [207]. Like any other design technique, the specific set of design processes and activities to use depends on the context in which the design approach is deployed. What “meaningful” stakeholder participation looks like in a given application context is guided by questions around *who* specifically should be involved, for *which tasks* within the measurement lifecycle, and in *what form*—and these decisions depend upon the goals of both the stakeholders who are participating, as well as the facilitators of the design process (e.g., HCI researchers, AI practitioners). Echoing recommendations from prior literature on participation and participatory AI (e.g., [98, 207]), these decisions should be made collaboratively with stakeholders whenever possible. If collaborative decision-making at the planning phase is not possible, these decisions should at least be made visible to stakeholders, which can empower them to contribute more meaningfully and critically. In this process, our framework offers a guide to systematically identify and discuss how specific properties of a given participatory measurement approach can constrain or enhance the extent to which stakeholders are truly able to leverage their expertise to inform measurement design. These decisions should be deliberated and documented prior to beginning a participatory engagement.

Relatedly, while we believe that scoring “higher” on most of the dimensions in the framework (with the exception of scalability¹) may be required to support true collaboration between AI expert and non-AI-expert stakeholders, where stakeholders are engaged as active designers and decision-making power is shared [112], this may not be possible or desirable in all contexts [438]. For example, *comprehensive* stakeholder participation across the Measurement Design Lifecycle may not always be achievable. In some industry contexts, AI development teams themselves may not have the power or flexibility to shape certain project parameters (e.g., whether to measure in the first place, or what specific concepts should be measured). In such cases, it may be best to be transparent about this with stakeholders, rather than asking them to spend much time on design decisions they cannot influence. Similarly, depending on the context, stakeholders may not have the capacity or interest to engage deeply across the entire measurement design lifecycle [438]. Even when participation is constrained due to context-specific constraints, our framework can help identify opportunities to support more collaborative approaches to participatory measurement within these constraints, focusing on dimensions where there are greater degrees of flexibility. Throughout a participatory measurement design process, researchers and practitioners should actively revisit questions around how to maximize benefits to stakeholders while minimizing risks and harms to their well-being. For instance, even when well-intended, researchers and practitioners who are not cognizant about the impacts that participatory approaches have on stakeholders may be susceptible to over-burdening them (e.g., through asking for too much time,

¹While we do not view *scalability* of participatory methods as being necessary to support more collaborative AI measurement processes in general, we note that it can be an important factor in project contexts that *require* large-scale input in order to meaningfully shape the design of a measurement instrument. For instance, the use of less scalable participatory methods in such contexts can sometimes threaten the *directness* with which stakeholders are able contribute, as developers are left to make decisions about how to compensate (e.g., by using synthetic data as a proxy for participant responses).

effort, emotional labor) [106] and can fall into pitfalls such as “participation washing” [383]. Researchers and practitioners engaging in participatory measurement design should actively strategize how to mitigate these risks and consider what *immediate* benefits participation can offer stakeholders [106], such as facilitating mutual learning and community building through the design process [420].

In the context of supporting participation for measurement of AI systems, several unique considerations discussed in prior literature on participatory AI apply. For example, it is unclear whether it is possible to support meaningful stakeholder participation for the design or evaluation of general purpose models that intend to provide broad capabilities applicable to multiple stakeholder communities with different values and goals [98, 411]. Instead, stakeholders may be meaningfully engaged in shaping measurement instruments (e.g., metrics, benchmarks, classifiers) to design or evaluate task-specific applications of LLMs [411]. As discussed in Opportunity 3, we also see opportunities to design future methods that help offset trade-offs between expressive and scalable participation (e.g., collection of inputs from larger numbers of stakeholders) through AI-assisted facilitation and analysis approaches.

Finally, some AI evaluation practices—while not considered “participatory” based on our inclusion criteria—involve soliciting stakeholder inputs to inform the measurement of AI systems. For instance, researchers in machine learning and natural language processing have explored various approaches to reinforcement learning from human feedback (RLHF), including variations in how human expertise is solicited and integrated to inform model improvements (e.g., [457]). Our framework’s dimensions (e.g., expressiveness, iterativeness, collectiveness) can also be used to systematically identify different ways in which the design of an approach can impact stakeholders’ ability to express their underlying beliefs, intents, and expertise, and also offer implications for improving such approaches. For instance, in typical RLHF practices, stakeholders are granted limited expressiveness when providing their perspectives and are asked to provide their perspective by choosing among pre-determined options. Even when finer-grained feedback is solicited from stakeholders, in current methods stakeholders are typically limited to annotating cases using small sets of pre-defined labels. They are also typically asked to provide labels individually, rather than having opportunities to discuss with others and consider alternative perspectives. Beyond the contexts of HCI and RAI, we hope that our framework can guide the development of future approaches in machine learning and natural language processing that make better and richer use of stakeholders’ expertise.

7.8.1 Limitations

Our paper has limitations which speak to additional areas for future research. First, our goal was to bring lessons from a multi-disciplinary literature review to the HCI and RAI communities. Our corpus is not necessarily comprehensive, and our analysis does not intend to speak to the entire space of approaches supporting stakeholder participation in measurement. While we included a breadth of potential keywords related to participation and measurement, there may be disciplines and papers using different terminology that our search did not capture. Relatedly, the dimensions of participation, synthesized based on our analysis, are not intended to be exhaustive. We invite future work to build upon these properties, to further enrich an understanding of how to support truly collaborative approaches to measurement design. Additionally, our paper reviews the methodological landscape of approaches in the

research literature, but does not examine how often given methods are used in practice. While out of scope for this work, we invite future research to examine the real-world prevalence of different methodological approaches.

7.9 Conclusion

With increased interest in supporting participatory approaches to AI design, alongside a growing call to center measurement issues in AI, there is an exciting opportunity to explore the intersection between these two efforts. In this paper, we draw on a rich multi-disciplinary literature, from the social and life sciences to philosophy of science, to learn from existing approaches to stakeholder participation in measurement. Based on our analysis of existing approaches, we present a methodological framework of properties that can hinder or facilitate a truly collaborative approach to participation in measurement. We conclude with a discussion on future research opportunities to guide emerging efforts toward effective stakeholder participation in AI measurement design.

Improving AI Measurement, Part 2

Exploratory Case Study on Worker-driven AI Measurement

This chapter is taken from the following paper:

Anna Kawakami, Chloe Zhao, Renee Shelby, Fernando Diaz, Haiyi Zhu, and Kenneth Holstein. “‘I want to be pushed, I want to grow’: Enabling social workers to design evaluations of LLM augmentation in their work.” In *Proceedings of the 2026 AAAI/ACM Conference on AI, Ethics, and Society*.

Abstract. Workers are increasingly asked to adopt AI systems to assist their work, yet are rarely given a voice in defining what meaningful AI augmentation should look like or how to evaluate for it. In this paper, we propose *worker-driven AI measurement*—a bottom-up approach to AI evaluation where workers collaboratively shape decisions about which tasks AI should augment, what “successful” augmentation looks like, and how it should be measured. We explore how to support this through a case study with 19 workers from a local school social work organization. Through a series of eight workshops, workers iteratively develop their own measurement goals for AI evaluation, systematize these goals, and then design a benchmark to capture how effectively an LLM can “challenge” them to reflect on their own assumptions and biases in the context of their day-to-day work. Workers collaboratively design and refine an LLM-as-a-judge rubric based on their professional and lived expertise. In validations of the worker-created benchmark, we find that there is strong agreement between worker and LLM judge ratings and that the resulting benchmark can differentiate performance across six state-of-the-art LLMs. Based on our case study, we discuss opportunities for future work to support worker-driven AI measurement as a complementary approach to existing top-down AI evaluation approaches.

7.10 Introduction

Frontline workers across occupations—from teaching and social work, to healthcare and finance—are increasingly asked to adopt AI systems to assist their work [44]. Yet, workers themselves rarely have a voice in defining what it would look like for AI to positively augment their day-to-day work, and how to design and evaluate for it [223]. Today, decisions about which parts of workers’ work to augment, how to augment them, and how to measure what “good performance” looks like are often left to technical experts or organizational leaders. Without worker involvement in these decisions, AI evaluations often reflect impoverished understandings of workers’ day-to-day work and the forms of expertise involved. For example, in occupations like social work and teaching, a lack of worker involvement has often led to AI deployments that conflict with worker values and practices [223].

A growing body of work focuses on developing AI evaluations for specific occupations [48]. For example, NLP and ML researchers have proposed new AI measurement instruments, such as benchmarks and metrics to measure the *automatability* of specific work tasks in healthcare, software engineering, finance, and design (e.g., [28, 440, 446]). More recently, researchers have begun to recognize the need to involve workers themselves in shaping AI evaluations [15, 28, 135]. Research in NLP and ML has explored methods to solicit feedback from workers to inform the design of AI evaluations. However, this work has typically constrained worker involvement to narrow sub-activities such as developing test cases or providing expert labels for benchmarks, without engaging workers in larger decisions such as what is worth measuring in the first place [224]. By contrast, HCI research has solicited workers’ perspectives on how new LLM-based systems should augment their professions in the future [e.g., 38, 454], but has not directly engaged them in the design of concrete measurement instruments. **Taken together, it remains an open question how to meaningfully support workers in the end-to-end, collaborative design of AI measurement instruments**—from *defining* what positive AI augmentation looks like from their perspective, to *operationalizing* practical measurement instruments that can accurately capture these notions.

In this paper, we present a case study with a local school social work organization that is actively exploring how to integrate LLMs into their work. In partnership with workers at this organization, we explore an approach to *worker-driven AI measurement design*—where workers collaboratively shape decisions around *which tasks* AI should augment, *what* “successful” augmentation looks like for those tasks, and *how* it should be measured. In particular, through an iterative series of eight workshops with 19 workers, we explore how to engage workers in (1) identifying realistic LLM use cases for which workers actually want LLM-based support, (2) developing a shared understanding of what “good” or “bad” LLM performance looks like for a given use case, and (3) operationalizing this understanding through the design of a benchmark (consisting of an LLM-as-a-judge system and a set of test cases).

Overall, through these design workshops, workers collaboratively developed an LLM-as-a-judge system to evaluate model performance on a specific use case core to their work—developing useful reflection questions to guide future meetings with teachers, based on workers’ observations of teacher practices and classroom challenges. Workers first collaboratively designed 16 realistic test cases that they believed would be challenging for an LLM, based on their past experience using LLMs for this use case. In early workshops, workers identified a measurement goal that is largely absent in existing industry AI evaluations: measuring how effectively LLMs

can function as “**people challengers**” (a play on current LLM tendencies to be “people pleasers.”) In the context of their work, this meant challenging workers to “*reflect on their own possible assumptions and biases [and helping] them expand their perspectives, providing opportunities for learning and growth.*”

Workers iteratively operationalized this goal into an LLM-as-judge rubric that captures their notions of what it looks like to meaningfully augment their work in their specific organizational and work context. Our validation shows that the resulting LLM-as-a-judge system has strong agreement with worker judgments, and that the benchmark can differentiate the performance of different state-of-the-art LLMs. Overall, we make the following contributions:

1. We propose *worker-driven AI measurement*—a bottom-up, collaborative design approach that complements existing evaluation practices for AI in work.
2. We explore this approach through an early **case study** with a local social work organization.
3. We present **opportunities for future work** to support worker-driven AI measurement design across worker and organizational contexts.

7.11 Background

In this section, we overview past work that brings social science measurement concepts to AI evaluation (Section 7.11.1). We then discuss nascent efforts to engage workers in shaping AI evaluations (Section 7.11.2).

7.11.1 AI Evaluation from a Social Science Measurement Perspective

Following recent calls in responsible computing and machine learning, in this paper we understand AI evaluation as a *social science measurement challenge* [5, 202, 203, 224, 441]. AI evaluations necessarily rely on assumptions about both *what* is important to measure and *how* unobserved, latent constructs (e.g., “fairness,” “accuracy,” “safety”) should be systematized and operationalized into practical measurement instruments, such as metrics and benchmarks. These assumptions often clash with reality in the downstream contexts where AI applications are used [202]. For example, developers and deployers make claims about the “utility” of healthcare LLMs using benchmarks based on medical exams [e.g., 316, 445, 462] that primarily measure an LLM’s ability to *recall* medical information [335]. However many of the tasks LLMs are used for in real-world clinical settings rely on other capabilities, and do not necessarily involve the recollection of medical information [335]. In fact, researchers have found that the majority of LLM benchmarks in the literature have poor construct validity [e.g., 26]. Others note that existing LLM benchmarks and metrics tend to jump from identifying broad measurement goals (e.g., measuring “harmfulness”) to operationalization, without articulating or critically interrogating how the underlying concept should be defined in the first place [441].

While prior work has begun to offer shared language and frameworks to ground AI evaluation in practices from social science measurement theory, comparatively little work has focused on how to meaningfully involve non-technical stakeholders in the design of measurement instruments for evaluating LLMs. Recently, researchers have drawn from ideas from measurement theory to explore how to engage non-AI-expert stakeholders in shaping LLM evaluations [203, 224, 387]—for

example, by engaging impacted communities in systematizing the concept of “cultural appropriateness” for evaluating text-to-image models [203].

To date, past work has explored how to support stakeholder participation in specific, isolated activities within the measurement design lifecycle. In this paper, we build upon this work by exploring how to support end-to-end stakeholder participation across the AI measurement design lifecycle—enabling stakeholders to collaboratively *identify* what concepts are meaningful to measure in the first place, *systematize* shared understandings of these concepts, and *operationalize* these systematized concepts into concrete AI measurement instruments.

7.11.2 Supporting Worker Participation in AI Measurement and Evaluation

Researchers evaluating AI systems for work increasingly recognize the need to involve workers themselves in shaping AI measurement instruments [15, 28, 135]. In recent years, the NLP and ML research communities have begun to solicit feedback from workers when developing benchmarks. For instance, researchers consulted clinicians to validate tasks used in MedHELM, a benchmark that evaluates LLMs on a taxonomy of five categories of medical tasks, going beyond existing benchmarks based on medical exams [28]. Similarly, HealthBench from OpenAI involved clinicians in creating test cases, writing rubrics for categories of health tasks, and providing labels for evaluation [15]. While these efforts demonstrate important steps towards meaningful engagement with workers in AI measurement, these works—like most existing evaluations of LLMs for work—focus on assessing how well LLMs can *autonomously* complete tasks rather than how well they can *support human workers’ activities*, despite LLMs being widely deployed to *assist* human work. Moreover, prior work has typically constrained worker involvement to narrow, isolated sub-activities (e.g., designing test cases or defining criteria), rather than engaging workers in the messy process of iterating across steps of the measurement design lifecycle—from defining measurement goals to systematizing and operationalizing latent concepts into measurement instruments. Past efforts have also typically focused on soliciting inputs from workers individually, rather than supporting collaboration, discussion, and co-learning among workers as is common in co-design approaches [224].

Meanwhile, the HCI research community has historically focused more on understanding workers’ broader visions of what the future of work should look like [38, 227, 454]. For instance, recent research has explored workers’ perspectives on how new LLM-based systems should augment their professions in the future [e.g., 38]. Other research has explored alternative futures of work, yielding broad implications for how to better measure and evaluate AI systems in the future [222, 395]. While research in HCI has more often engaged workers in collective discussions and design activities, these efforts have primarily focused on surfacing workers’ broader visions and perspectives. To our knowledge, no prior work has explored how to directly engage workers in *co-designing AI measurement instruments* that can help capture progress towards these visions.

Our paper presents a first exploration of how to support workers in the end-to-end, collaborative design of AI measurement instruments—from *defining* what positive AI augmentation looks like from their perspective, to *operationalizing* practical measurement instruments (i.e., a benchmark) that can accurately capture these notions.

7.12 Methods

To understand how to support workers in designing a measurement instrument to evaluate LLMs in their work, we formed a collaboration with a school social work organization based in Pennsylvania, United States. At the time of this research, the organization employs around thirty social workers who work with local preschool, kindergarten, and elementary schools, often in low socio-economic status communities. The workers conduct observations of teachers' classrooms, reflect on challenges and opportunities observed, then meet with teachers to collaboratively discuss paths forward. Overall, workers' goals are to better bring out teachers' existing strengths so that the teachers, in turn, can better support the children's growth and wellbeing.

In March 2025, the organization began exploring the use of LLMs to support workers' practices. The director of the organization provided workers with an option to use a customized LLM, developed using OpenAI's ChatGPT customization service using a high-level set of system instructions specifying desirable model behaviors (see Appendix for details). Given that the organization lacked context-specific evaluations of their LLM, the director was very interested in designing new evaluations to understand how well their organization's LLM actually worked with hopes to identify opportunities to improve it or create better ones in the future.

Over the course of eight weeks, we conducted an iterative series of eight workshops with 19 workers across the organization to explore how different design activities and scaffolds can support worker participation in benchmark design. In this section, we describe our exploratory approach to supporting worker participation (Section 7.12.1), workers' background (Section 7.12.2), our analysis approach (Section 7.12.3), and our validation of the resulting benchmark (Section 7.12.4).

7.12.1 Exploratory Approach and Activities

To support the collaborative design of an LLM benchmark, we explored an approach that builds upon ideas from measurement theory in the quantitative social sciences [5, 8, 201] and design [54, 98, 112, 365]. We break down AI evaluation design into three core phases, and explore how we can support meaningful worker participation in each: (1) Identifying meaningful LLM use cases to focus on, (2) Developing a shared understanding of workers' use cases and measurement goals, and (3) Operationalizing measurement goals into an implementable measurement instrument (a benchmark consisting of an LLM-as-a-judge rubric and a set of test cases).

To support participation, we created design activities guided by the following principles:

- **Guiding principle 1: Support expressive participation.** Design activities that allow workers to fully express their perspectives to shape foundational decisions in measurement design (e.g., what to measure and how), without constraining their perspectives into pre-defined categories. Existing approaches to engaging stakeholders in AI measurement design, particularly in the operationalization phase, tend to constrain workers' abilities to express their ideas [224].
- **Guiding principle 2: Support informed design iteration.** Ensure that workers are empowered to assess the consequences of design decisions—for example, by inspecting how their design decisions affect the behavior of a measurement instrument on realistic data—to inform ideation and refinement. Existing approaches to engaging stakeholders in measurement design often do

not provide stakeholders with opportunities to meaningfully assess the implications of their design decisions, let alone to iteratively refine their design decisions [224].

- **Guiding principle 3: Support meaningful collaboration amongst workers.** Supporting forms of design ideation and assessment that allow workers to learn from and build upon each others' perspectives is particularly important in organizational and group settings, where AI deployments can have collective impact. Yet, most approaches to supporting stakeholder engagement in measurement design solicit input from individual workers, with minimal opportunities for co-learning and collaboration among stakeholders [224].

In the following, we briefly describe the three phases.

Phase 1: Collaboratively Exploring Potential AI Use Cases to Focus On.

This phase focused on exploring which LLM use cases to focus on for benchmark design. After identifying a use case, we then had workers design a set of test cases that represent realistic ways in which they prompt GPT in their own work. Overall, we conducted a series of activities with 11 workers:

Understanding the breadth of workers' current and desired uses for LLMs (pre-workshop activities). Prior to the first workshop, we asked all workers at the organization to asynchronously share ways they were currently using LLMs to assist their work and ways they might like to use future, improved LLMs to assist their work. This resulted in 39 current and 18 desired uses written by nine workers. Then, to better understand their responses, we conducted 30 minute formative interviews with five of the workers. During each interview, we asked clarifying questions about their use cases, how they perceived the performance of their organization's current model on the use cases, and how they actually prompted the model for a given use case. Whenever possible, we asked workers to illustrate with concrete scenarios and examples from their past interaction logs.

Developing a shared understanding of workers' LLM use cases & identifying an initial one to focus on (workshop 1). Based on workers' documentation of use cases and formative interviews, we identified an initial set of six use cases that represent different ways in which workers were (trying to) use the model to augment their work. We then conducted a 90-minute workshop with eight workers. During this workshop, workers discussed, refined, and elaborated upon the six use cases. Specifically, we aimed to understand (1) how many workers desired LLM assistance (or were already using LLMs) for a given use case, (2) how workers currently prompted their existing model for a given use case, and (3) whether the initial set of use cases had any major gaps. Based on workers' discussions, our research team identified a use case that was both high priority for workers, and where workers hoped to see improvements in LLM performance. Workers were interested in using LLMs to generate reflective questions that could usefully challenge their thinking and assumptions about their classroom observations. We include more details on the use case selection criteria in Appendix D.1.1.

Designing an initial set of test cases (workshop 2). At the start of workshop 2, we gave workers an overview of the goals and practices for designing test cases.

Drawing on workers' background in educational settings, we asked them to imagine that they were designing tests that all future AI models must take in order to prove their ability to effectively help workers with the use case. We encouraged workers to think about test case scenarios that they think may challenge the model's capabilities, drawing upon their existing experiences trying to use LLMs for this use case. We additionally encouraged workers to think about realistic test case scenarios where they would actually find it valuable to receive LLM-based support. Workers developed an initial set of test cases in small groups (3 - 4 individuals per group). As workers designed their test cases, workshop facilitators circulated around the room to provide light guidance when needed (e.g., about key missing context they might want to include in their prompts). Through this process, workers collectively designed a shared set of 16 test cases, which then served as a foundation for the subsequent workshops. We lightly augmented this initial set of test cases to support rubric refinement in Phase 3, as described in Appendix D.1.5.

Phase 2: Identifying Measurement Goals & Systematization.

During this second phase, our goal was to help workers develop a shared understanding of what "good" versus "bad" performance looks like for their use case. To support this, we had workers inspect and discuss pairs of model responses to their test cases. Through iterative discussion across multiple response examples, workers began to collaboratively identify high-level aspects that capture important dimensions of variation in model response quality. Overall, this phase included two workshops with 13 workers:

Annotating (un)desirable properties of model responses (workshops 3 - 4). At the beginning of workshop 3, we reminded workers to keep their standards high and grade based on how they believe an ideal future AI system should respond in given cases. We emphasized to workers that *they* are the experts in how they want AI to augment their work. During this workshop, workers worked in small groups to inspect examples of model responses on the test cases they had designed in the previous workshop. For each test case, workers saw three model responses. To increase variation among model responses—with the goal of helping workers discover relevant dimensions of difference—we generated each model response based on a random selection of base models (from OpenAI, Anthropic, Google) and system instructions, as described in Appendix D.1.3.

In their small groups, workers examined the two pairs of model responses for a test case they were randomly assigned. Workers then discussed and took notes on what they thought was working well in each model response versus what they thought could be improved. Workshop facilitators circulated around the room, listened in on conversations, and provided light guidance as needed (e.g., clarifying the task we're asking them to do, asking workers reflective questions to support their thinking). After workers identified several high-level areas for model improvement within their group, we asked each group to share their findings with the rest of the workshop attendees. We asked workers to phrase these in a few brief sentences following the template "Responses should ..." We encouraged workers to build upon, challenge, or ask clarifying questions about each others' ideas. Workshop facilitators took notes on the identified areas for improvement using a shared online board (Figma) that was viewable to everyone in the workshop. Through collaborative discussion, the researchers and workers synthesized the shared ideas into a small set

of high-level goals for measurement, each linked back to workers' discussions and annotated examples of model responses.

For the next phase, we focused on an important area for model improvement that workers identified: LLMs' ability to effectively challenge workers to think more deeply and critically, rather than just taking workers' initial characterizations of a classroom situation at face value.

Phase 3: Collaboratively Operationalizing a Measurement Goal into a Practical Measurement Instrument.

During this phase, we focused on supporting workers in operationalizing this concept of effective "people challenging" into an LLM-as-a-judge rubric. To support worker participation in this phase, we conducted four workshops with 10 participants. During the first workshop of this phase, we introduced "best practices" for designing rubrics, drawing on existing resources online [e.g., 351]).

Designing an initial rubric (workshop 5). Workers designed an initial rubric in small groups, using our provided rubric template. The rubric template encouraged participants to break down their identified high-level measurement goal into a more specific set of rubric criteria, and to provide concrete positive and negative examples for each criterion. To inform the design of the rubric, we asked workers to consider their past discussions and response annotations around their "ideal" model response, as well as past test cases they had annotated in previous workshop activities, which we handed back to workers. After the workshop, our research team further refined the rubric that workers created. Our refinements were informed by workers' discussions of the criterion, drawn from transcriptions and notes of prior workshop activities. For example, we added rubric sub-criteria that were not represented in the existing design of the rubric but discussed in previous workshops. We additionally noted follow-up questions to ask participants at the next workshop, and left placeholders for places we were uncertain what the workers' intent might be. We validated all refinements made at the start of the next workshop, and workers either rejected or substantially modified several of these suggested refinements.

Iteratively assessing and improving the rubric design based on feedback (workshops 6 - 8). In the final two workshops, workers iteratively refined the rubric design by examining how well an LLM judge using their rubric aligned with their own assessment of given model responses. In particular, workers were shown two model responses to each test case (generated using the same approach to the systematization phase, as described in Appendix D.1.3), then were asked to "score" each model response according to the individual criteria in their rubric. Workers then inspected how an LLM scored the same model response using their rubric, identifying criteria where the LLM score misaligned with their score. Workers discussed potential underlying reasons for misalignment by looking for potential ambiguities in the original wording of their rubric criteria—with the option to view an LLM-generated rationale for the LLM score to aid in this task—and then cross-checking these against the specific test case prompt and model response they were evaluating. Workers then directly revised the design of their assertion and reran the LLM to see if the rubric revision brought the LLM's score and rationale into alignment with their own judgment. If it did not, workers continued to explore other rubric revisions with the goal of aligning the LLM score with their own. Workers repeated this process iteratively, across multiple test cases and model responses.

7.12.2 Participant Information

In total, 19 workers (around 2/3 of the organization) participated in the workshop. Some workers participated in only one workshop, while others attended most of the workshops. Appendix D.1.6 (Table D.2) includes an overview of number participants per workshop and Table D.1 includes participant demographics. Participants were compensated \$15 for every hour of the workshop they participated in. The director of the organization also compensated them, so they received their hourly wage through workshop engagement.

7.12.3 Analysis Approach

To analyze both the process and outcomes of the participatory workflow, we analyzed a combination of materials: Approximately 12.75 hours of workshop transcripts, an initial and final version of the worker-developed rubric, and the test cases. To analyze the workshop transcripts, we used an abductive analysis approach guided by specific research questions for each phase of the workshop. We developed the guiding questions through team discussion to ensure our analysis captures both outcome-oriented and process-oriented insights. Following a shared open coding session to calibrate on coding granularity, two researchers qualitatively coded each workshop transcript. For each workshop, we ensured that at least one of the researchers coding the workshop transcript was the workshop facilitator. While coding, the first researcher aimed to capture observations that directly address the guiding question for that workshop phase, while also remaining open to coding other observations. The second coder validated the first researcher's codes by marking areas of disagreement or coding additional observations not captured by the first coder. The two coders then resolved disagreements through synchronous discussion. We conducted a bottom-up affinity diagramming process to iteratively refine and group the resulting codes into successively higher-level themes.

7.12.4 Benchmark Validation

We conducted two validations of the benchmark workers developed. First, we conducted a validation survey to compare workers' judgements with an LLM judge's assessments of model responses. We first removed four test cases that were used to refine the rubric in Phase 3 (test cases 7, 9, 14, 16). For each of the remaining 12 test cases that were not used in rubric refinement, we generated two model responses. We used the same strategy to generate model responses from Phase 2 and 3, as described in Appendix D.1.3. For the LLM judge, we used the same GPT-5 model and system instructions we used for the rubric refinement activity. 12 workers participated in the validation activity. All 12 workers attended at least one prior workshop to contribute to the rubric design. Two workers were assigned to independently respond to one model response. For each model response, each worker saw the test case, model response, and the LLM judge score. They were asked whether they agree or disagree with the LLM judge score, and to provide their own score if they disagree. Using this dataset, we analyzed worker-judge agreement, including both worker1-judge and worker2-judge agreement (accounting for the two individual worker responses) and the average worker-judge agreement (where we compare the LLM judge response to the average of the two workers' responses).

For the second validation, we examined how well the benchmark differentiates across model performance for six different state-of-the-art LLMs: Gemini 3.1 Pro, GPT-5.4 pro, Claude Opus 4.6, Gemini 3 Flash, GPT 4o, and Claude Sonnet 4.6. We

specifically selected the most recent and advanced models from Google, OpenAI, and Anthropic at the time of the analysis (the first three models above) as well as their most widely used models (the last three). For each model, we generated responses using a minimal system instruction (i.e., “You are a helpful AI assistant for a school social work organization.”) To compare how different models perform when given more specific information about task requirements, we also generated responses using a tailored system instruction drawn from the rubric that workers designed. In our analyses, we compared how the six LLMs across the two system instruction conditions perform on individual criteria in the rubric. See Appendix D.1.3 for details.

7.12.5 Limitations and Ethical Considerations

Our case study is intended to be an early-stage exploration of worker-driven AI measurement. We acknowledge that our case study, while deeply engaging with workers in an organization, is limited in scope and generalizability. The organization we partnered with is small (around 30 employees), and the exploratory workflow we used in this case study may not generalize well to larger organizations. That said, the lessons learned from this focused case study provide valuable insight into where worker participation is critical across the three measurement phases, and how participation might be supported across organizational contexts. The Discussion section provides implications and opportunities for future work.

We additionally acknowledge the limitations of our validation approach and pre-post survey. For example, our study does not evaluate whether AI systems that score high on workers’ benchmark *actually* better support workers in doing their work better. Future work should explore ways to evaluate how well the LLM-as-a-judge system actually anticipates the impacts of an AI system on workers’ performance, for example, by tracing the downstream outcomes of workers’ AI-assisted decisions. Moreover, our validation is limited in scope due to the small number of test cases used. Future work could support larger-scaled validation, for example, by using worker-designed test cases to seed a larger set of test cases that can be used for validation. Additionally, while the pre-post survey was taken by the majority of the participants who regularly engaged in the workshops, we cannot make any causal claims on the pre-post survey findings (Findings Section 7.13.5) due to the small size of the organization and number of participants.

Finally, our study was approved by our institution’s IRB, and we received explicit consent from workers before their participation. We ensured that workers’ test cases are fully anonymized (e.g., not naming any specific individuals or institutions). We acknowledge that our workshop engagements were labor intensive, and we are grateful to workers at the organization for sharing their time, expertise, and curiosity about the research with us. In the Discussion, we reflect on opportunities to reduce the labor- and time-intensiveness of future approaches without trading off workers’ abilities to express their perspectives.

7.13 Designing a Benchmark with Workers

Through early workshops, workers identified effective “**people challenging**” as an important measurement goal—i.e., measuring how well LLMs can move beyond “people pleasing” tendencies, and instead “challenge the writer to reflect on their own possible assumptions and biases [and] help them expand their perspectives,

providing opportunities for learning and growth.” Workers then iteratively designed an LLM-as-a-judge system that includes **ten distinct criteria**, each of which captures a distinct dimension of how workers operationalize this concept in the context of their work.

In this section, we present key findings from the benchmark design process, organized sequentially from earlier- to later-stage design activities. First, we discuss **(1) how workers designed challenging test cases** for an LLM to handle effectively, which helped ground the overall rubric design process in rich, realistic scenarios (Section 7.13.1). By discussing examples of model responses to these test cases, **(2) workers identify measurement goals for LLM evaluations that are largely absent from existing industry evaluations**—for example, assessing how well LLMs challenge workers’ assumptions and beliefs, to support them in reflecting more deeply and growing as human experts (Section 7.13.2). Moreover, through an iterative process of discovery, **(3) workers designed, assessed, and refined a ten-criterion rubric—capturing important nuances and edge cases that go beyond their initial systematization**, by drawing on their professional and lived expertise during operationalization (Section 7.13.3). Through our validations, we find that the resulting worker-designed **(4) LLM-as-a-judge system strongly aligns with worker judgments** on the test cases and that the resulting **(5) benchmark can differentiate model performance across state-of-the-art LLMs** (Section 7.13.4). Finally, beyond the benchmark produced, we found that **(5) workers may experience other benefits from engaging in the workshop**, like learning more about AI capabilities and limitations.

7.13.1 Designing realistic test cases

After identifying a use case to focus on (Methods Section 7.12.1), workers designed a set of test cases for their use case. Overall, when instructed to design test cases that may challenge models’ capabilities, **workers designed a set of cases that convey rich, nuanced, and diverse scenarios inspired by their day-to-day classroom observations**. For the remainder of the design process, these test cases served an essential role in shaping workers’ discussions about what measurement goals to prioritize and how to define them (Section 7.13.2), as well as how to operationalize measurement goals into an LLM rubric. While we briefly elaborate on the content of the test cases in this section, Appendix D.3.1 includes complete test case examples.

Workers built on their direct experiences working in classroom settings to generate a range of complex scenarios where they might turn to an LLM for help. For example, they described scenarios where a child is struggling to express their needs (Test case 1), a teacher inappropriately using humor when talking to a child (Test case 5), and a child is saying racial slurs to a new teacher (Test case 9). To help the LLM interpret these observations, workers intentionally include a range of contextual details about these scenarios—like the environmental setting (e.g., race, gender, or age of children involved, socio-economic status of the school region) and temporal details (e.g., changes in teacher or child behaviors over time). Test cases also often mention teachers’ personalities and qualities (e.g., “She is warm, but firm and has high expectations”, Test case 14).

Because workers see LLMs as a tool for aiding self-reflection rather than prescription, they are also intentional about including their own assumptions and direct reactions from their observations. For example, workers include their subjective interpretations of events (e.g., “the child [...] feels unseen and misunderstood by his teachers”, Test case 1). Others describe their personal reactions they had as an

observer of a situation (e.g., “Watching this whole interaction, I was struggling to regulate myself, and feeling angry on behalf of the child...”, Test case 4). As Sections 7.13.2 and 7.13.3 later describes, the presence of subjective interpretations and personal feelings in these test cases helped workers surface and articulate their measurement goals (e.g., to assess how well an LLM challenges them in reflecting on their assumptions and biases in their work, so that they can best learn and grow as human workers).

7.13.2 Identifying and defining meaningful measurement goals

Overall, workers identified measurement goals that are largely absent from existing industry evaluations—in particular, the need for models to challenge workers’ assumptions, promote critical thinking, or avoid behaviors that may conflict with their learning and growth as human workers. While there is growing discourse on the need to deploy AI systems that enhance worker expertise, the current landscape of benchmarks offer minimal resources for actually measuring these properties. In the following, we first describe how workers’ past experiences using their organization’s current LLM shaped their desires for improved LLM performance. We then discuss how workers began to articulate shared high-level ideas for what to evaluate in model responses, through carefully examining and comparing examples of model responses to their test cases.

In the past, workers had tried—but often failed to—prompt the LLM to serve as a reflective aid, rather than a problem-fixer.

For most use cases (e.g., preparing for future teacher meetings), workers desired an LLM that could serve as a *reflective aid* rather than as a source of prescriptive advice. However, they recalled prior experiences struggling to effectively prompt the LLM to support this goal. Even with explicit prompting, they observed that the LLM tended to be overly prescriptive. For example, when prompting it to support reflection around a concerning interaction between a teacher and a child, workers described experiences where it instead provided solutions to position the worker as a “hero” who independently resolves the concern. Workers expressed concerns around how this orientation conflicts with the nature of their work and values behind it. As one worker voiced, “[we] are not fixers ... [we] are wonderers.” Workers described a need for the LLM to respect and support their intentions to stay “curious” about classroom occurrences, rather than approaching challenges with a mindset to find immediate solutions.

By discussing examples of model responses to test cases, workers were able to more concretely articulate what effective “people challenging” looks like in their context.

These discussions led workers to articulate several distinct high-level dimensions of model response quality, as well as the concept of an LLM being an effective “people challenger” in contrast to a “people pleaser.” For example, in early workshop discussions, one worker summarized: “I want to be pushed, I want to grow. I want to take perspectives that I can’t take by myself. Because that’s the whole point of using the tool.” By examining examples of current model responses on their test cases, workers articulated a range of potential improvements in this direction. For example, one group of workers voiced that they want LLM responses to “prompt the writer to dive deeper

[into the context] for additional information” on other stakeholder perspectives, rather than assuming that the prompter has provided the full story. Workers in this group articulated this desire, after observing an example of a test case where the writer included reflections on one actor but not others. Workers also voiced a desire to have the LLM challenge the writer to reflect on potential biases in their prompts. As one worker voiced: “Depending on our background, we might be more likely to [elevate] the child’s voice or more likely to raise the teacher’s voice or the parent’s voice, right, in a particular prompt, and we want it to notice that and promote further exploration.”

7.13.3 Iteratively designing and assessing an LLM-as-a-judge rubric

In this section, we discuss how workers iteratively operationalized the concept of “people challenging” in the context of generating reflective questions (Section 7.13.2), into a comprehensive rubric with ten distinct criteria. Workers began with the following statement, along with a set of annotated examples of model responses from their prior workshops:

“The response should not be a people pleaser. Instead, it should challenge the writer to reflect on their own possible assumptions and biases. It should also help them expand their perspectives, providing opportunities for learning and growth.”

Building off this statement and their prior discussions, workers developed an initial rubric with five criteria, then iteratively refined it into a more comprehensive version with ten criteria through discussion and observations of how an actual LLM judge interpreted their rubric. Table 7.2 contains an overview of descriptive counts comparing the initial rubric and final rubric at the start and end of the operationalization phase. Appendix D.3 includes the complete initial and final versions of the rubric.

Overall, workers operationalize “people challenging” in ways that differ from related industry efforts. While the concept is related to measuring sycophantic behaviors [66, 185], existing sycophancy benchmarks focus on undesirable behaviors to *avoid*, whereas workers prioritized defining what *good* responses should look like in their specific work context. By drawing upon their professional and lived expertise grounded in their day-to-day work, and examining examples of how an LLM judge uses their rubric to assess response quality, workers iteratively reflect upon and discover important nuances in how they conceptualize different criteria.

Counts of properties	Initial Rubric	Final Rubric
Words (characters)	324 (2036)	1490 (9505)
Criteria	5	10
Conditional state- ments	0	4
Edge cases	0	8
Elaborations reflecting professional expertise or local knowledge	1	10

TABLE 7.2: Descriptive counts comparing the initial versus final rubric, at the start and end of the operationalization phase where workers iteratively refined their rubric.

Workers primarily focus on designing criteria to measure progress towards positive visions rather than just to avoid undesired behaviors

We found that, in breaking down their high-level goals for measuring LLM behaviors into concrete criteria, workers primarily focus on positive criteria to capture the concept of “people challenging” (criteria 1 - 7). A comparatively smaller subset (criteria 9 and 10) are negative criteria, focusing on behaviors to avoid. In particular, the ten criteria workers designed span four main categories: (1) **Expanding workers’ perspectives** around the people and situations surrounding them (criteria 1, 2, 3, and 4), (2) **Promoting critical self-reflection** on their assumptions or potential biases (criteria 5, 6, and 7), (3) **Avoiding people pleasing tendencies** like giving superficial compliments (criteria 8 and 9), and (4) **Providing external resources** for further learning (criterion 10). Table 7.3 provides an overview of the criteria in each of these four categories.

Rubric categories & criteria
<i>Category 1: Expanding workers’ perspectives around the people and situations surrounding them</i> Criterion 1: Always challenge the writer to consider all people whose perspectives might be relevant, beyond the ones they mentioned in their prompt. Criterion 2: Always challenge the writer to consider all possible perspectives of the people mentioned. Criterion 3: Challenge the writer to consider other interpretations of people’s actions, if the writer’s prompt includes an assumption about someone’s intent and motivations. Criterion 4: Challenge the writer to reflect on their role in the situation as a component of the environment, if they have not already.
<i>Category 2: Promoting critical self-reflection on their assumptions or potential biases</i> Criterion 5: If the writer does not mention their own assumptions or interpretations, challenge them to do so. Criterion 6: Challenge the writer to reflect or elaborate on their personal experience of the situation, regardless of whether they mention it. Criterion 7: Identify potential biases in their interpretation of the situation.
<i>Category 3: Avoiding people-pleasing tendencies</i> Criterion 8: Don’t people please or provide superficial compliments. Criterion 9: Don’t over-center the role of the facilitator.
<i>Category 4: Providing external resources for further learning</i> Criterion 10: Challenge the writer to explore additional resources to inform their thinking.

TABLE 7.3: High-level representation of the final rubric contents, including four main categories synthesized from the criteria. See Appendix D.3, Table D.4 for the full rubric.

Workers began with an initial five-criterion rubric based on their systematization, and expanded it into a more elaborate rubric through iterative testing and refinement

Overall, by seeing how well an LLM judge using their rubric aligned with their own assessment of given model responses, workers expanded their initial rubric into a final rubric with substantially more elaborations, concrete examples, and edge cases. The final rubric contains two times more criteria (10 criteria versus 5 criteria) and four times more words (324 words versus 1490 words) compared to the initial rubric. As the next subsection describes, some of the new criteria surfaced as workers realized their existing criterion conflated multiple distinct ideas and should be broken down into distinct criteria. In other cases, researchers added new initial criteria after noticing that workers had discussed preferences for model behaviors that were not already captured in their list of criteria; workers then refined these criteria to ensure they accurately represented their perspectives.

Additionally, whereas the initial rubric implicitly assumed all criteria should always apply regardless of the user's prompt, the final rubric specifies in what cases a given criterion should apply (e.g., stating the model should *always* challenge the writer to reflect on their role versus specifying the model should only do this *if* the writer has not done so already (criterion 4), and elaborating on what these cases may look like). Seeing how an actual LLM judge rates specific model input-output pairs helped workers reflect more deeply on the conditions *when* a given model response is appropriate based on a written prompt. In other words, the task of refining the LLM judge's assessments of model behavior prompted workers to further unpack the factors underlying their own assessments. In the Appendix, Table 7.2 includes a high-level comparison of the initial versus final rubric.

Through the process of operationalization, workers capture important nuances and edge cases beyond their initial concept systematization, drawing on their professional and lived expertise

We observed how, through an iterative process of discovery during the operationalization phase, workers were pushed to reflect more upon their values and beliefs around what good model behaviors look like, and to more clearly articulate these by refining criteria. Workers drew upon their own professional expertise as front-line social workers, including their knowledge of different kinds of school contexts, to inform rubric refinements—including when making sense of why an LLM judge using their rubric may not have replicated their own assessment on a given criteria. For example, workers engaged in rich discussions with one another to collaboratively align on the boundaries of what counts as “potential biases” in their specific work. While bias is a commonly operationalized concept in AI evaluation, by seeing how an LLM judge assessed potential biases in model responses to their test cases, workers began to operationalize the concept in ways that diverge from standard AI evaluations. At the start of operationalization, criteria 7, on prompting the worker to reflect on any potential biases, originally started as a two-sentence description (39 words); through discussion, workers iteratively expanded the criteria description to include concrete edge cases and examples grounded in their specific work contexts (288 words). For example, reflecting on examples of model responses, workers explained how even seemingly neutral comments like “the classroom is very loud” may reflect the writer's biases because their background influences their interpretation of the volume of the classroom. The workshop facilitators asked workers to

include a corresponding example of a statement that is *not* culturally biased to ensure the boundary will be understood by the LLM judge, leading to the addition: “The teacher stated that they are overwhelmed by the volume of the classroom.” In another example, a worker noticed that a test case described a child being racist towards a teacher and asked the LLM for reflective questions to help improve the relationship between the two individuals. After seeing that the LLM judge did not interpret this as a potential bias in the writer’s prompt, the worker reflected that suggesting an *individual* relational approach to address a broader systemic issue may signal a bias in the writer’s interpretation. The worker expanded the criterion to include this as an edge case.

7.13.4 Rubric Validation

We report on two validations of the benchmark that workers designed (described in Methods Section 7.12.4).

To what extent does the LLM judge align with worker judgments?

Overall, we found a 96% correlation between worker and judge ratings across the set of 12 held-out test cases that workers did not use during rubric refinement. Worker ratings were computed as the average of the two worker ratings on a given case. We find that worker-judge agreement is higher than the agreement between the two individual workers assigned to a given case (worker1-worker2 agreement), suggesting that the judge’s rating may represent an “average” perspective among workers. This may be because the rubric itself was created through a collaborative, consensus-based process. Table 7.4 contains the worker-AI and worker1-worker 2 correlations across the ten criteria.

Criteria No.	Worker1-Worker2	AI-Avg Worker	AI-Worker1	AI-Worker2
1	0.749	0.888	0.853	0.808
2	0.811	0.856	0.763	0.871
3	0.991	0.990	0.996	0.979
4	0.842	0.987	0.872	0.976
5	0.536	0.942	0.613	0.988
6	0.847	0.958	0.892	0.951
7	0.569	0.470	0.494	0.347
8	0.988	0.997	0.988	1.000
9	0.859	0.978	0.966	0.918
10	0.682	0.917	0.768	0.935

TABLE 7.4: Results of the first validation activity showing Pearson’s correlation for worker-AI outputs. The cells are color coded as follows: $x \geq 0.9$ is green, $0.9 > x \geq 0.7$ is light green, $0.7 > x \geq 0.4$ is light red, $0.4 > x$ is red.

This is especially apparent in cases where there may be greater disagreement among workers about how to interpret a given criterion. For example, criterion 5

specifies “If the writer does not mention their own assumptions or interpretations, challenge them to do so.” As workers pointed out during the workshop discussions, workers may disagree with one another about whether a given statement is an “assumption” or “interpretation”. For criterion 5, the agreement among workers is weaker (0.536) whereas the average worker-judge agreement is stronger (0.942). The same holds for criteria 4, 6, 9, and 10.

Notably, both worker-AI correlation (0.470) and worker1-worker2 correlation (0.569) are low for criterion 7, which asks the LLM judge to “identify potential biases in their interpretation of the situation.” As workers pointed out during the workshop discussions, workers may disagree with one another about whether a given statement is potentially “biased.” Beyond that, however, the criterion may be particularly challenging to operationalize effectively. It is possible that, during the rubric refinement phase, the operationalization of criterion 7 overfitted to the specific test cases used for refinement. For example, criterion 7 may currently capture aspects of bias that are represented in the four test cases used for refinement, while failing to capture the broader underlying concept.

Model	Custom	Minimal	Δ
claude-sonnet-4-6	3.898	2.959	+0.939
claude-opus-4-6	4.487	2.858	+1.629
gemini-3-flash-preview	4.348	2.791	+1.557
gemini-3.1-pro-preview	4.777	2.736	+2.041
gpt-4o	2.996	2.581	+0.415
gpt-5.4-pro	4.600	2.901	+1.699

TABLE 7.5: Average scores (out of five) on workers’ LLM-as-a-judge system, for each base model and system instruction condition (customized and minimal system instructions). Delta is difference in average scores between the conditions.

Can the benchmark differentiate among state-of-the-art base models?

We found that workers’ benchmark can differentiate among different state-of-the-art LLMs. We found very strong correlations across three iterations of generating model responses (average pairwise correlation 0.911), so our findings here report on the average across the three generations. Overall, we find that all six LLMs (without tailored system instructions) perform poorly on rubric criteria that measure *positive* LLM behaviors, while performing better on criteria that capture *negative* behaviors to avoid (criteria 9 and 10). An exception is criterion 7, which measures how well an LLM identifies potential biases in the writer’s interpretation of a described situation. We additionally find that using a custom system instruction based on the worker-created rubric improves the performance of all six LLMs for most rubric criteria, but still leaves room for improvement. Finally, we find that, when using customized system instructions, the three *most recent and advanced* models outperform the three *most commonly used* models. There are no major differences across models when the minimal system instructions are used. Table 7.5 shows aggregate scores across criteria, separated by base model and system instruction condition. Figure 7.2 and 7.3 shows a heatmap of scores by criterion (y axis) and base model (x axis), colored by

mean score. Figure 7.2 includes scores for the minimal system instructions condition, while Figure 7.3 includes scores for the customized system instruction condition.

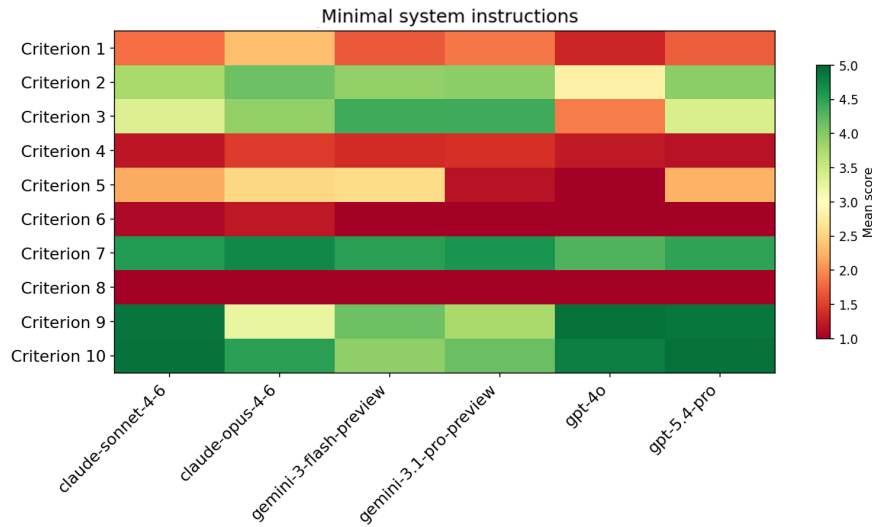


FIGURE 7.2: Heatmap showing the average scores per criterion for each base model using the minimal system instructions.

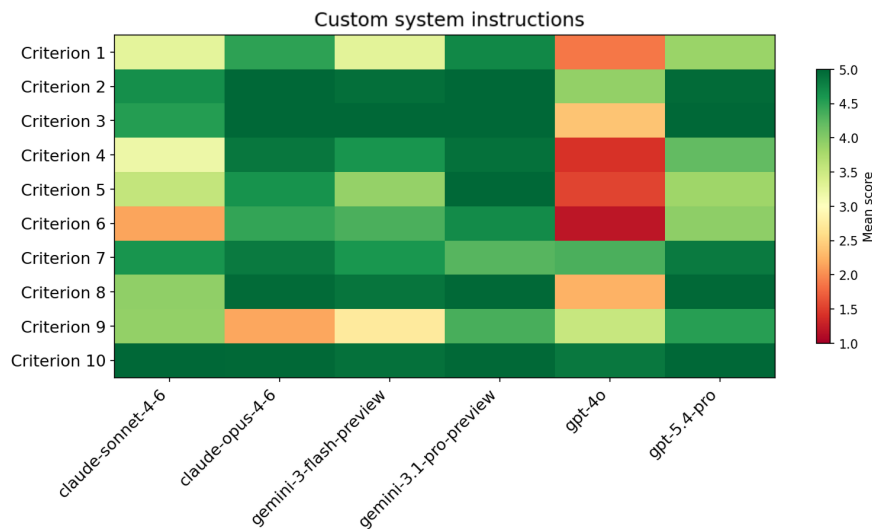


FIGURE 7.3: Heatmap showing the average scores per criterion for each base model using the custom system instructions.

7.13.5 Workers' Experiences from Participation

Based on the pre-post survey, we found several attitude shifts following workers' engagement with the workshops. First, on average, workers' agreement that AI produces correct information decreased slightly from pre- to post-workshop ($MD = -0.30, n = 10$), indicating that workers may have become more critical of AI performance after engaging in the workshop compared to before, even though they had used AI in their work previously. Second, workers also reported reduced concern about AI replacing their own work skills ($MD = -0.40$) and about AI-driven employee replacement more broadly ($MD = -0.50$); for both these survey items, five of ten participants indicated lower concern after participating in the workshops. During the workshops, workers repeatedly described how the design activities (e.g.,

ideating measurement goals, refining the rubric) helped them reflect on how they might make better use of LLMs in the future, by improving their understanding of how their prompt design may impact LLM model responses. Appendix D.2.2 includes additional details on workers' experiences. We discuss implications of these findings in the Discussion.

7.14 Discussion

In our study, workers wished to measure a concept for which benchmarks do not currently exist: how effectively a model can *challenge* them to think and reflect critically in their day-to-day work. While this concept, at first glance, appears closely related to the concept of "social sycophancy" [66], social sycophancy benchmarks focus on measuring how well LLMs *avoid undesirable behaviors*. By contrast, workers' concept of "people challenging" represents a positive vision, capturing *desirable* model behaviors, beyond the mere absence of sycophantic behavior.

Furthermore, workers' goals were not to measure "people challenging" in the abstract, but rather, to measure what "people challenging" means in the context of their specific work and organizational context. Because of this, the workers' direct involvement in systematization and operationalization activities, not just concept identification, was essential. Only workers themselves would understand the nuances of their concept well enough to operationalize it into criteria that speak to their particular needs and desires. For example, in our study, workers' rubric refinements were informed by cultural differences they experience between their own backgrounds and their observed classroom settings (which were often in low SES communities), as well as their own specific values that shape and are shaped by their organizational culture (e.g., desiring reflective over prescriptive responses).

Therefore, *worker-driven AI measurement design*—where measurement instruments are designed by workers themselves through a bottom-up, collaborative approach—can play an important *complementary* role to existing paradigms of AI evaluation. While current paradigms typically aim to develop AI measurement instruments that apply at broader levels of abstraction and generality, a more bottom-up approach is needed to better reflect what matters to workers across a diverse range of occupations, worker roles, and organizational contexts. Indeed, as our case study shows, a worker-driven approach can lead to identifying entire measurement goals that are otherwise unrepresented among existing AI measurement instruments. In practice, we envision that worker-driven measurement design efforts may be initiated at different levels, from the "micro", hyper-local level (e.g., individual organizations like our partner in this study) to the "meso" level (e.g., coalitions of workers that engage in similar forms of work across organizations, within a broader occupation). Efforts at either of these levels may create a starting point for adaptation to other work contexts. For example, a measurement instrument designed by workers in one organization could be adapted for use in similar organizations, reducing the need for each organization to create their own bespoke instruments from scratch. Relatedly, future work could also explore how to design for a *plurality* of perspectives in worker-driven AI measurement, enabling workers to customize their own AI measurement instruments to represent shared high-level concepts. By designing to preserve meaningful disagreement [fazelpour2025value], worker-driven AI measurement approaches can support a wider range of uses, from enabling individual advocacy to developing multi-faceted shared understandings of model performance.

We also observe preliminary signs that worker-driven AI measurement processes may provide additional benefits that go beyond outcome-oriented benefits like the resulting AI measurement instruments (Findings Section 7.13.5). For example, after participating in our workshops, workers expressed greater recognition of the value of their work as humans and were more *critical* of the performance of AI systems. While the current study cannot support causal claims about why these attitude shifts occur (see Limitations section 7.12.5), future work should further investigate such potential benefits of worker-driven AI measurement approaches.

Our case study surfaced several additional opportunities for future approaches to best make use of workers' expertise and time [106]. For example, future work could explore approaches to automatically and immediately signal areas of potential improvement in workers' rubrics. For instance, future interfaces might show workers real-time previews of how specific edits may impact downstream LLM judge behavior, while they are iteratively refining their rubric. Future approaches can also explore how to help workers more systematically identify edge cases that their rubric should account for. For example, future interfaces might automatically generate new test cases to stress test the rubric design, in order to support worker reflection toward rubric refinements [246]. Importantly, our findings suggest that, even with such process enhancements, workers themselves must continue to drive design and refinement decisions, drawing on their unique expertise to reflect upon, evaluate, and shape LLM judge behavior.

To summarize, our exploratory case study offers the following implications for practitioners and researchers interested in designing evaluations of AI-augmented work:

- To design and evaluate AI systems that *augment* worker skills, engage workers themselves in identifying specific measurement goals and concepts. These measurement goals should supplement rather than replace existing measurement goals that evaluate other AI capabilities that may be overlooked by workers but surfaced by other stakeholders (e.g., in our case study, workers paid less attention to measuring sycophancy).
- Beyond identifying measurement goals and concepts, engage workers in both the *systematization* and *operationalization* stages of measurement design. Workers have unique expertise that complement the technical expertise required to develop AI measurement instruments for specific application contexts.
- Once AI measurement instruments are developed, publicly share the measurement instrument. Sharing the instrument with other application contexts (e.g., with similar measurement goals, AI use cases, and work contexts) can support broader adaptation and use.
- Worker-driven AI measurement design could potentially additionally be used to support critical AI literacy for workers.

7.15 Conclusion

We propose *worker-driven AI measurement design* as a complementary approach to existing AI evaluation practices, where workers collaboratively shape decisions about what “successful” augmentation looks like in the context of their work and how it should be measured. In an exploratory case study with a local school social work organization, workers developed a benchmark (an LLM-as-a-judge system and a set

of test cases) to operationalize what it means for an LLM to meaningfully *challenge* their thinking within their day-to-day work. We find that workers are able to draw on their professional and lived expertise, grounded in their specific organizational context, to identify, define, and operationalize measurement goals beyond the space of existing AI measurement instruments. Overall, this case study suggests promise for future approaches aimed at empowering workers to drive the design of AI measurement instruments and evaluations.

Chapter 8

Concluding Reflections

Over the course of my PhD, my research was driven by a desire to advance more positive futures for AI and work—where AI systems meaningfully augment workers' expertise and capabilities rather than diminish them. I've focused specifically on AI in socially complex work, and within that, feminized occupations (where worker expertise is historically devalued and misunderstood). This is where I felt most drawn to making an impact. From the very start of my PhD, I conceptualized research impact broadly: not just as advancing the research field, but as improving real-world practices and informing policy for AI and work. Towards achieving this broader notion of impact, I've formed close collaborations with workers and organizations to understand and intervene on challenges, engaged with policymakers on my research, and written policy-facing artifacts that translate from research to policy.

Having completed my PhD dissertation, there is still so much I want to do. The work that I currently feel the most dissatisfied about (and the greatest urge to continue pushing) centers translating knowledge and interventions developed from research to real-world practice. For example, I aspire to do research that develops tools that can actually be adopted and used across multiple organizational contexts, and that can, in turn, inform new guidelines and standards for how best to develop and evaluate AI systems for work contexts. Too often, promising interventions are born from research projects and then remain in papers, shortcutting their potential to shift real-world practice.

So, what's next? First, I became a US citizen, around one year ago in January 2025. While I didn't feel ready to give up my Japanese citizenship, I moved forward with my application knowing I would one day want to serve in the federal government. Studying AI in real-world organizations, often in state and local government, I felt again and again that many of the challenges I observed cannot be addressed through academic research or technical innovations alone. Instead, they needed to be coupled with broader structural changes (e.g., developing guidelines and standards, setting field-wide priorities, supporting cross-sector collaboration) that shift incentives and infrastructural resources. These types of changes can often best be achieved through work at the federal level. By becoming a US citizen, I wanted to give myself the opportunity to address these challenges directly.

In the next stage of my life, I will be an AI research scientist at the Center for AI Standards and Innovation (CAISI), historically known as the "AI safety institute" of US government. CAISI is a small AI evaluation-focused organization based within NIST at the Department of Commerce in Washington, DC. Here, I'll continue to push forward on these open challenges—doing research that advances not just knowledge and prototypes, but practical interventions and guidelines that help shift toward a world where AI systems meaningfully uplift workers, organizations, and society more broadly.

Appendix A

Worker-AI Interaction, Part 2

This chapter includes the appendix for Chapter 2 on the following publication:

Hao-Fei Cheng*, Logan Stapleton*, **Anna Kawakami**, Venkatesh Sivaraman, Yanghui Di Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. “How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions.” In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*.

A.1 Numbers and percentages of children by race and risk level

See Table A.1 for a breakdown of the numbers and percentages of children referred to CYF for discretionary referrals by risk level (High, Medium, Low, and mandatory screen-in).

A.2 Appendix terminology

In addition to screen-in rate disparities and decision accuracy, we also evaluate the following additional metrics.

	All children	Black children	white children
All risk levels	51750	26123 (50.5%)	21623 (41.8%)
Mandatory screen in	15161 (29.3%)	9639 (36.9%)	4863 (22.5%)
High risk	31022 (59.9%)	18536 (71.0%)	11013 (50.9%)
Medium risk	11778 (22.8%)	5208 (19.9%)	5653 (26.1%)
Low risk	8950 (17.3%)	2379 (9.1%)	4957 (22.9%)

TABLE A.1: Numbers and proportions of children by race and AFST risk level. Percentages are relative to the total number of children per race. For example, the 9639 Black children labeled mandatory screen in made up 36.9% of all 26123 Black children referred to CYF from August 2016 to May 2018.

- Precision: the percentage of children that are screened-in and are either removed from home within 2 years, or re-referred again within 2 months of the referral.
- True Positive Rate (TPR): the likelihood that a child who has the positive proxy ground truth label (i.e. either removed from home within 2 years or re-referred again within 2 months of a referral) will be screened-in.
- False Positive Rate (FPR): the likelihood that a child with a negative proxy ground truth label (i.e. neither removed from home within 2 years nor re-referred again within 2 months of a referral) will be screened-in.

A.3 Comparisons of other disparity metrics and thresholds

Figure A.1 shows racial disparities in screen-in, accuracy, true positive, false positive, and precision rates across possible AFST-only thresholds from pre- to post-AFST deployment. Figure A.2 shows screen-in, accuracy, true positive, false positive, and precision rates by race, possible AFST-only thresholds, and AFST-only versus worker-AFST decisions from August 2016 to May 2018.

A.4 Workers see risk labels as screening recommendations

Neither official AFST documentation nor public comments from CYF leadership claim that a Low risk label (scores 1 through 9) means that a referral should be screened out nor that a High risk label (scores 15 and up) means screen in. However, this is the message that is implied through the interface design of the AFST and is what we heard from workers. For one, the AFST interface shows scores binned into Low, Medium, and High risk levels with clear divisions between them. Low risk is green (calm, low stress) and High risk is red (urgent, severe). Although not explicitly stated in the AFST documentation, the message that Low risk referrals should be screened out and High risk referrals screened in is relayed to workers via these design choices. When we interviewed and observed workers, many implied or echoed this explicitly. One caseworker (erroneously) said that *“with a High risk [referral], you know, we absolutely have to screen them in.”*¹ Another said, *“when it is High risk, I’ve just been going with open investigation,”* i.e. screen in. While observing one caseworker, they got a referral and explained: *“it was a Low risk [referral]... And instead of screening out, I just recommended accept for investigation.”* Here, the caseworker implied that Low risk referrals should be screened out. Furthermore, five caseworkers and one supervisor said that they did not know what Medium risk scores meant or did not find them helpful when making screening decisions. In all, workers see an AFST Low risk label as a recommendation or pressure to screen out a referral, High risk label to screen in, and Medium risk confers no recommendation.

A.5 Accuracy breakdown by score

In this section, we examine the accuracy of AFST-only decisions (at the threshold of 15) versus worker-AFST decisions for referrals broken down by AFST score or mandatory screen-in label from August 2016 to May 2018. Accuracy is calculated

¹Workers are not mandated to screen in High risk referrals. However, this quote exemplifies workers’ perceived pressure to screen in High risk referrals.

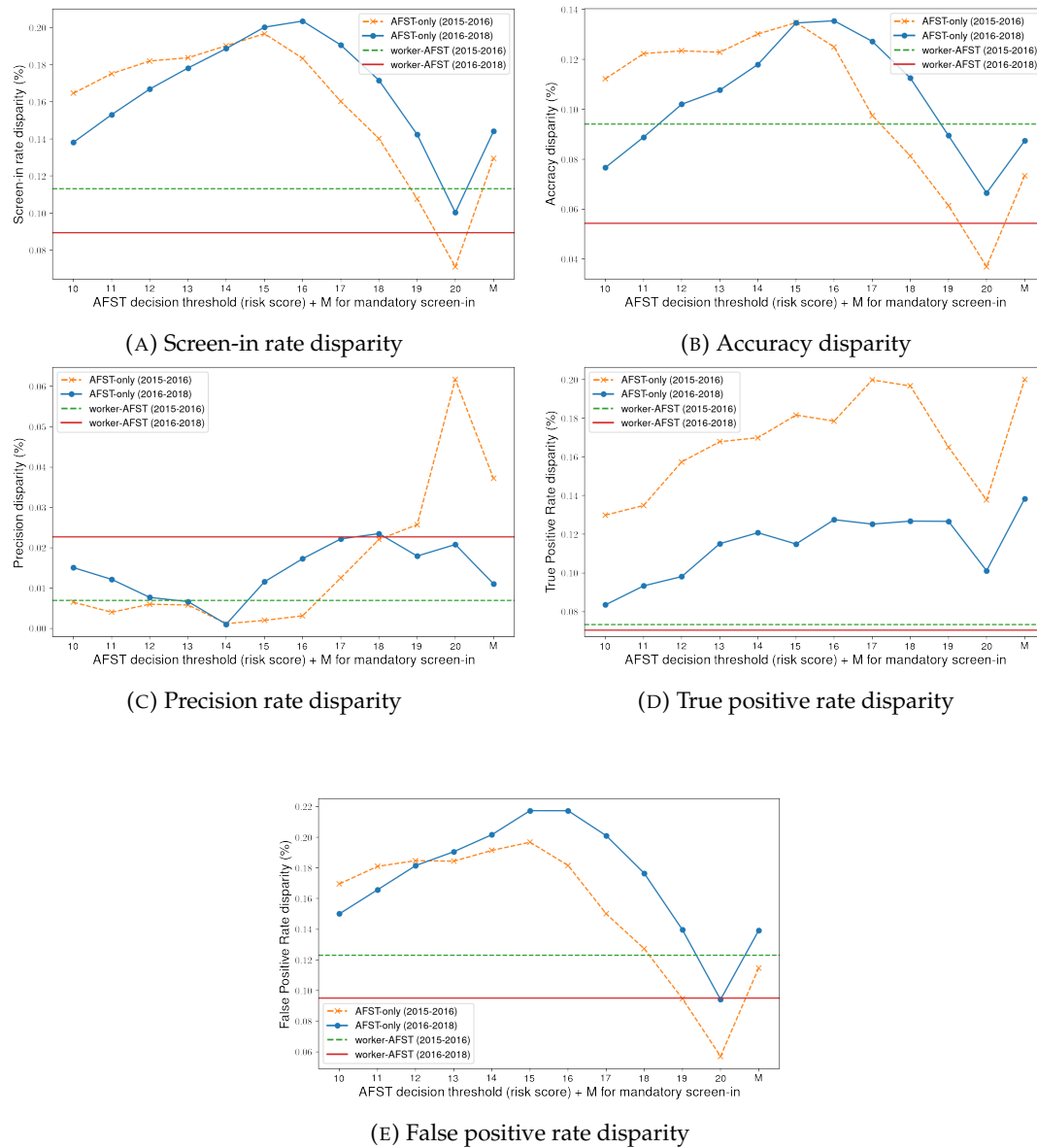


FIGURE A.1: Common group fairness metrics showing disparities between Black and white children. The x-axis is the various decision thresholds for the hypothetical AFST-only decisions (for a given threshold, AFST-only would screen in all referrals with that score or above, and screen out all referrals under). The solid red line represents the worker-AFST final decisions after AFST deployment (2016-2018), the solid blue line with dots represents AFST-only decisions in the same time period (2016-2018). The dashed green line represents the worker-only decisions without before the AFST was deployed (2015-2016), and the dashed orange line with crosses represents the AFST-only decisions generated retrospectively for all referrals in the same time period (2015-2016).

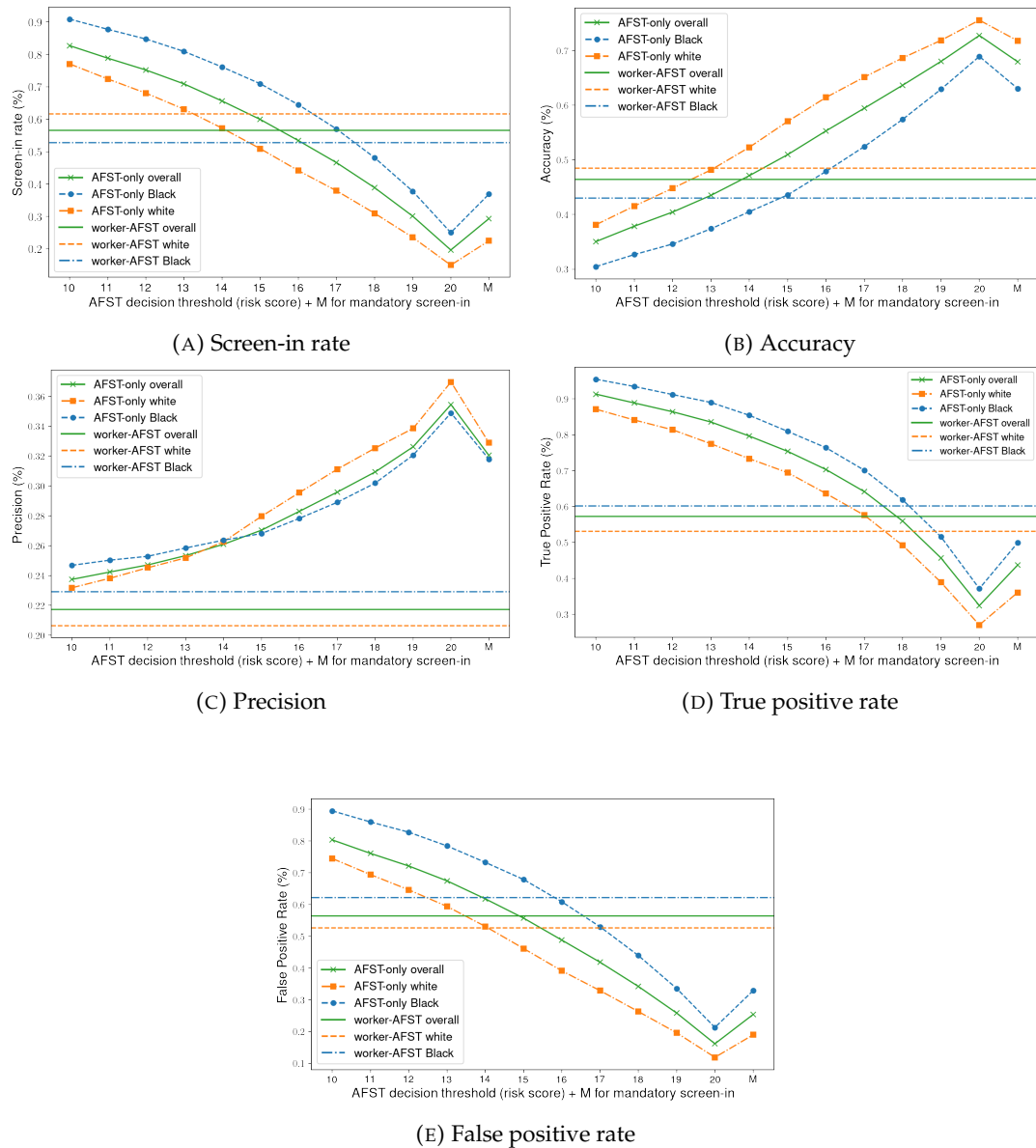


FIGURE A.2: Comparisons of decision outcomes between worker-AFST and AFST-only decisions for referrals between 2016 to 2018. The x-axis is the various decision thresholds for the hypothetical AFST-only decisions (for a given threshold, AFST-only would screen in all referrals with that score or above, and screen out all referrals under).

AFST score	Num. of children by score	AFST-only accuracy	Worker-AFST accuracy
15	3374	16.7%	48.1%
16	3518	19.4%	40.8%
17	3998	22.8%	40.3%
18	4532	25.2%	38.7%
19	5396	27.3%	39.3%
20	10169	35.4%	39.3%
M	15161	32.1%	37.6%

TABLE A.2: Accuracy rates AFST-only and worker-AFST policies binned by AFST score or mandatory screen-in (M) from August 2016 to May 2018 for High risk referrals. The AFST-only policy (with a threshold of 15) would have screened in all referrals in this table, so the accuracy rates in the third column are equivalent to the percent of children who were either re-referred within 2 months or placed within 2 years. For example, of the 15161 children labeled mandatory screen-in (placement model score 18 or higher), 32.1% were re-referred or placed in foster care. The rightmost column is the accuracy of workers' actual screening decisions.

just as in Section 2.13.1, except that the denominators are only referrals given a single AFST score. For example, if all 15,161 children labeled as mandatory screen-ins were screened in (as the AFST-only policy would have), then 32.1% of these decisions would have been accurate. The actual screening decisions made by workers were accurate for 37.6% of these 15161 children. Table A.2 shows that worker-AFST decisions were more accurate than AFST-only decisions would have been for every score from 15 and above. Table A.3 compares the accuracy of AFST-only decisions versus worker-AFST decisions for Low risk referrals (with scores 1 to 9) from August 2016 to May 2018. Worker-AFST decisions were less accurate than AFST-only decisions for every score from 1 to 9. Overall, these results suggest that worker discretion increased the accuracy for children that the AFST-only would have screened in and decreased it for those the AFST-only would have screened out. In particular, worker discretion increased the accuracy of decisions made in mandatory screen-in referrals, which are recommended to be screened in by default. However, these results are again limited by problems with evaluating counterfactual predictions, as discussed in Section 2.11.

AFST score	Num. of children by score	AFST-only accuracy	Worker-AFST accuracy
1	81	87.7%	84.0%
2	329	94.2%	79.3%
3	619	93.7%	71.9%
4	805	90.8%	64.2%
5	1063	90.1%	60.6%
6	1243	87.5%	61.5%
7	1363	89.7%	62.0%
8	1587	88.3%	57.8%
9	1847	87.1%	60.5%

TABLE A.3: Accuracy rates of AFST-only and worker-AFST policies binned by AFST score and mandatory screen-in (M) from August 2016 to May 2018 for Low risk referrals. Any AFST-only with threshold 10 or above would have screened out all referrals in this table, so the accuracy rates in the first column are equivalent to the percent of children who were neither re-referred within 2 months nor placed within 2 years. For example, of the 81 children with an AFST score of 1, 87.7% were neither re-referred nor placed in foster care. The right-most column is the accuracy of workers' actual screening decisions.

Appendix B

Improving AI Governance

This chapter includes the appendix for Chapter 6 on the following publication:

Anna Kawakami, Amanda Coston, Haiyi Zhu, Hoda Heidari, and Kenneth Holstein. “The Situate AI Guidebook: Co-Designing a Toolkit to Support Multi-Stakeholder, Early-stage Deliberations Around Public Sector AI Proposals.” In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.

B.1 Deliberation Questions

This section includes the full list of questions included in the *Situate AI Guidebook*¹ (Ver.1.0).

B.1.1 Goals and Intended Use

The set of questions below are intended to support conversations around the following broader question: **Given our underlying goals and intended use case(s), is our proposed AI tool appropriate?** This stage would benefit from the expertise of the following stakeholders at the minimum, amongst others: Agency leadership, AI practitioners, frontline workers, community members.

Overall goal for using algorithmic tool

Who the tool impacts and serves

- Who is going to be affected by the decision to use this hypothetical AI tool?
 - Who is going to be the most impacted?
- Who benefits from the use of the tool?
 - To what extent are the targeted outcomes intended to benefit the agency, versus the community?

Intended use

- What evidence do we have suggesting that the painpoint this tool aims to solve actually exists?

¹<https://annakawakami.github.io/situateAI-guidebook/>

- What evidence do we have suggesting that technology may offer a remedy to this painpoint? (Evidence may include, for example, historical agency metrics, legislature, community members, research reports.)
 - What evidence suggests the specific form of technology we are envisioning (e.g., predictive analytics) may offer a remedy?
- What are the additional challenges and risks associated with pursuing a technological solution to this problem?

Involving agency-external stakeholders in determining the goals

- Think about the most impacted stakeholders you identified in response to the questions above. How do we bring their voices to the table when determining goals?
- How can we open opportunities for those who are most impacted by the new tool to inform the decision-making process?
- When will we start to engage impacted communities in discussions around how the tool should be designed or used?

Differences in goals

- Are there differences in the goals the agency versus community members think the tool should address? If so, what are they? If we are uncertain, what can we do to understand potential differences?
- What evidence do we have that we adequately understand the outcomes the community cares about?
- To what extent are we optimizing the things the agency cares about versus what impacted community members care about?
- Is the process we have in mind for achieving a community-oriented outcome (e.g., child safety) also aligned with the community's desires?

Envisioned harms and benefits

- What are the potential harms and benefits of the tool, and to whom?
 - Do benefits outweigh the harms?
 - Do we expect there to be tradeoffs between accuracy, fairness, explainability? For example: making decisions in a completely random fashion may look "fair", but is not necessarily accurate.
 - Will this tool help us better allocate (scarce) resources?
- What biases (as a government agency) do we bring into this decision-making process?
 - How can we identify and mitigate them? What forms of collaboration (e.g., with impacted community members) can help us do this?
- How does this tool help us better deliver to the people we are serving, if at all?

Selection of outcomes that the algorithmic tool aims to improve**Impacts of outcome choice**

- Hypothetically, imagine that our tool does a perfect job of improving the outcome that it targets. What additional problems might this create elsewhere in the system?
- To what extent are we optimizing the things the agency cares about, versus what impacted community members care about?

Assumptions behind outcome choice

- What assumptions are we making, when deciding what the tool should optimize?
- How are we operationalizing goals for the tool, e.g., improving child 'safety'? What assumptions are we making?
- How do we bring providers to the table to decide on the use of outcomes?

Predictability of outcomes

- Have we run any tests on historical data records, to check whether we get predictions on this outcome that actually make sense?
- How rare is the event we are trying to predict? If it is rare, how reliably do we think we can predict it?
- How does the inclusion of additional information (e.g., attributes) improve outcomes?

Empirical evaluations of algorithmic tool**Measuring improvement based on outcomes**

- Once the tool is deployed and in use, how can we evaluate how well it is working in the short-term? How can we evaluate how well it is working longer-term?
- What are some ways we might evaluate whether this tool is successful in improving the targeted outcomes?
- For evaluating worker-ADS decisions post-deployment: Do the decisions change by worker experience, worker demographic, or by supervisor?
- What performance measures do we plan to use to evaluate the tool?
- What performance measures have already been used in early analyses of historical data, prior to the deployment of the tool?
- Does this tool improve outcomes? How are we operationalizing "improve"?
- How does the use of the tool compare with the status quo? E.g., can we demonstrate the tool improves outcomes for the population of interest?

- What is the "performance" and "fairness" of the existing baseline/status-quo decision process?
 - Is there someone with relevant domain expertise that could help explain anomalies or trends?
- Do we think there are tradeoffs between accuracy, fairness, explainability? If so, what are they?
- How are we measuring negative and positive impact on families?
- Is there someone with relevant domain expertise that could help explain anomalies or trends?
 - How well do you understand the domain application of the historical data used in evaluation?
 - Are there changes in policies and domain-specific practices in the historical data?
- Are there measured improvements resulting from the model's deployment?
- Are we using appropriate evaluation methods, e.g., synthetic controls, discontinuity analysis when cutoffs on risk exist.
- What outcome measures are we evaluating on? What can these measures tell us, and what can they not tell us?

Centering community needs

- How can we effectively evaluate the tool from the perspective of impacted community members?
 - E.g., what does false positive, false negative mean for different impacted communities? How are we weighting false positives and false negatives, in a given use case, based on the relative costs of each type of error for the impacted stakeholders?

Worker perceptions

- How might front-line workers respond to the tool? How can we better understand their underlying concerns and desires towards the tool?
- How do front-line workers perceive the algorithm? (e.g., do they consider it a top-down requirement or a useful tool)
- Do domain experts also believe the model 'makes sense', e.g., selection of important features?

B.1.2 Societal and Legal Considerations

The set of questions below are intended to support conversations around the following broader question: **Given the societal, ethical, and legal considerations and envisioned impacts associated with the use of AI tools for our stated goals (identified in Facet 1), is our proposed AI tool appropriate?** This stage would benefit from the expertise of the following stakeholders at the minimum, amongst others: AI practitioners, frontline workers, community members, legal experts.

Legal considerations around the use of algorithmic tool

- Do the people impacted by the tool have the power or ability to take legal recourse?
- Is there clarity around policies, e.g., whether algorithmic outcomes are included under 'public records'?
 - If someone asks for information around the tool, but there's no precedent, does the agency know what to do?
- Are you having conversations with the Department of Justice and attorneys, to make sure the new decision models you implement will follow existing policies, procedures, statutes, and rules?
 - Do you know which design decisions will be dictated by the law? For example: In the context of child maltreatment screening, if certain conditions are present in a case, then it is legally required to screen in for investigation.
- Can you inform existing policies, procedures, statutes, and rules to better meet the needs of new decision models?
- Do you need a new temporary rule to receive permission to use the model?
- How are you interpreting challenges to ambiguities in prior legal decisions around the use of the tool?
- What are challenges to interpreting legal documentation and guidelines?
 - How well can we interpret case-specific considerations in the context of legal documentation/guidelines (e.g., when there is a lot of grey in practice, but the law is written in black and white)?
 - * E.g., in child maltreatment: "threat of harm" or "physical abuse" allegation type sounds black/white but there are various factors that make this grey. E.g., how hard did it hit them? Did it leave a mark? Action occurred but no impact from the action?

Ethical and fairness considerations around the use of algorithmic tool**Impacted Community Member Needs**

- Are there differences in the goals the agency versus community members think the tool should address? If so, what are they? If you are uncertain, what are your plans for understanding potential differences?
 - What are the envisioned harms and intended benefits from the tool that impact the community and the agency?
- Can we have impacted community's representatives or advocates at the table, to inform the design and use of the tool?
- How well are we engaging people closest to the problem and those impacted through the entire design, development, implementation, maintenance process?

- Are the outcomes intended for agency or community benefit?
- How well do we understand what outcomes the community wants to improve?
- Do we understand how impacted stakeholders perceive each decision? E.g., emotional valence, potential impacts, etc.
- To what extent are we optimizing the things the agency cares about versus what impacted community members care about?

Involving Impacted Communities

- What are underlying assumptions that tool developers/researchers may have, regarding the soundness of the design decisions made in the tool?
- How can we set up external participation opportunities, to increase access?
 - E.g., avoiding scheduling during a 9-5pm period (to open involvement to those who want to be involved)
 - E.g., is it possible to involve groups that are not involved and paid by the agency, to get input and feedback?
 - Do we know who should be included? How can we build the right network of people to talk with?
- Who has a seat at the table, to decide how the tool impacts you?
- How are you engaging with people closest to the problem (e.g., frontline workers, community members, or others impacted by the decisions)?
- Have you communicated the limitations and historical context of the data, to community members?
- How well do we understand the costs, risks, and effort required of community members, if we invite them? E.g., many were directly harmed by decisions made by the agency.
- When do we start to engage impacted communities into discussions around the design or use of the tool?

Clarity of Ethics Goals and Definitions

- Can we agree on a definition of fairness and equity in this context? What would it look like if the desired state is achieved?

Operationalization of Ethics Goals

- Are fairness and equity definitions and operationalizations adequately context-specific? (For example, in the child welfare domain: children with similar profiles receive similar predictions irrespective of race?)
- Do we know how to appropriately operationalize our fairness formulation in the algorithm design?
- Can we mitigate biases in the model?

- How can we balance tradeoffs between false negatives and false positives?
- How well are we integrating domain-specific considerations into the design of the tool?
- Have we recognized and tried to adjust for implicit biases and discrimination inherent in these social systems that might get embedded into the algorithm?

Envisioning Potential Negative Impacts

- Do we understand the negative impacts of the decision made across sensitive demographic groups?
- What are the externalities / long-run consequences of the decisions?

Social and historical context surrounding the use of algorithmic tool

- Have we recognized and tried to adjust for implicit biases and discrimination inherent in these social systems that might get embedded into the algorithm?
- How might we clearly communicate the limitations and historical context of the data to community members?
- Are you modeling historical, systemic patterns?

B.1.3 Data and Modeling Constraints

The set of questions below are intended to support conversations around the following broader question: **Given the availability and condition of existing data sources, and our intended modeling approach, is our proposed AI tool appropriate?** This stage would benefit from the expertise of the following stakeholders at the minimum, amongst others: AI practitioners.

Understanding data quality

- How does the data quality and trends compare with an 'ideal' state of the world?
 - What does our data look like, in terms of different demographic outcomes?
- Has the definition of the data changed over time? (E.g., in child welfare, has reunification always meant to reunify with the parent?)
- What data do we have access to?
 - Do we have the data/feature set to replicate the tool/analysis/and predictive accuracy of the existing tool?
- How well do we understand the meaning and value of the data that will be used to train an algorithm?
- How is the quality of this data?
 - How accurate is the data?
 - How recent is the data?

- How relevant is the data?
- Has the data been consistently collected?

Process of preparing data

- How are we preprocessing the data?
- Who should be involved in making decisions around whether to include or exclude certain data points or features? Do we have plans for involving those people?
- How do we address bias in the data?
- Do we have metrics for feature importance, that we could show relevant domain experts?
- How well do we understand the data collection process?
- Data leakage questions: Are we preventing oversampling of certain populations?
 - E.g., in child welfare: Are we pulling one child per report, and one report per child, to ensure there's no information leakage between training and test sets?

Model selection

- Is our model appropriate given the available data? Why or why not?

B.1.4 Organizational Governance Factors

The set of questions below are intended to support conversations around the following broader question: **Given our plans for ensuring longer-term technical maintenance and policy-oriented governance, do we have adequate post-deployment support for our proposed AI tool?** This stage would benefit from the expertise of the following stakeholders at the minimum, amongst others: Agency leaders, AI practitioners, frontline workers.

Long-run maintenance of algorithmic tool

Measuring changes in model performance over time

- Do we expect there will be shifts in performance metrics over time? If so, why? What are our plans for identifying and mitigating those shifts?
- Do we expect that the data collection process will improve over time? What might this imply for how we maintain the tool? E.g., Is there a need for adjusting thresholds over time?

Mechanisms to identify long-run changes

- Are we repeating feature engineering efforts over time?
 - Are we detecting how trends shift over time at the population level?
- Are there mechanisms in place that track whether certain data features have changed over the years?
- Do we have mechanisms to track longer-term outcomes over time, so that we can monitor for changes in model performance?
- Do we have the mechanisms to monitor whether the tool is having unintended consequences?

Organizational policies and resources around the use of algorithmic tool**Policies around worker interactions**

- Is there training for frontline workers who will be asked to use the tool? What evidence suggests that this training is adequate?
- How are frontline workers trained?
- Is it clear to workers what information the tool can access, and what information it cannot?
 - How is this communicated to workers?

Governance structures

- Imagine that we could assemble the “ideal team” to monitor and govern the tool after it is deployed: What are the characteristics of this ideal team?
 - Who is the *actual* team that will monitor and govern the tool after it is deployed?
 - Given the gaps between the “ideal team” and the actual team we expect to have: What risks to post-deployment monitoring and governance can we anticipate? How might we mitigate these risks?
- Are there appropriate forms of governance, around the implementation?
 - Do those involved in governance have domain knowledge in the application context and have knowledge of the implementation process?
- Are there sufficient guardrails in place to ensure algorithms wouldn’t get weaponized?
 - E.g., IRB-like programs and researchers at the same table, to minimize risk of weaponizing?

Internal political considerations around the use of algorithmic tool

- How well do we understand system administrators' and leadership's perspectives around the use of this tool?
- How well do staff and leadership understand 'why' the tool could bring value?
- Do system administrators and leadership perceive this tool positively?
- Do leadership support the future use of the tool?
 - Do we have backing at a leadership level? E.g., director, agency, governor, community partners?
- Is there sufficient buy-in from middle managers and executive support?
- Do we have mechanisms to address concerns that could come up during the ideation and design process?

Appendix C

Improving AI Measurement, Part 1

This chapter includes the appendix for Chapter 7 on the following paper:

Anna Kawakami, Haiyi Zhu, and Kenneth Holstein. “AI Measurement as a Collaborative Design Practice: A Multidisciplinary Synthesis and Future Directions.” Under review.

C.1 Literature search queries

To develop our initial corpus, we searched digital libraries and Google Search for the following keywords within the title and abstract: {participat* OR stakeholder OR design* OR sketch* OR “domain expertise” OR “domain expert” OR expert OR elicit* OR “human feedback” OR align* OR operationaliz* OR involve* OR communit*} AND {measure* OR metric OR indicator OR proxy OR feature OR model* OR variable OR outcome}.

C.2 Papers excluded by criterion

In Figure C.1, we present a PRISMA diagram [moher2010preferred] that illustrates how we applied our four inclusion/exclusion criteria to create a final corpus.

C.3 Table of Papers

This section includes a table of the papers included in our corpus. For each paper, we include the paper number, title, phase of measurement design it covered, and reference.

PNum	Paper Title	Phase	Paper Link
P1	Quality Indicators of Pharmaceutical Care in Palestinian Integrative Healthcare Facilities: Findings of a Qualitative Study among Stakeholders	Phase 2	[377]

Continued on next page

(Continued from previous page)

PNum	Paper Title	Phase	Paper Link
P2	Consensus Development of Priority Outcome Domains for Community Mental Health Cares by Multiple Stakeholders: Protocol for an Online Delphi Study in Japan	Phase 1	[381]
P3	AINeedsPlanner: A Workbook to Support Effective Collaboration Between AI Experts and Clients	Phase 3	[231]
P4	Is a Seat at the Table Enough? Engaging Teachers and Students in Dataset Specification for ML in Education	Phase 3	[415]
P5	Sustainability Assessment in the German Detergent Industry: From Stakeholder Involvement to Sustainability Indicators	Phase 2	[373]
P6	Towards Intersectional Feminist and Participatory ML: A Case Study in Supporting Femicide Counterdata Collection	Phase 3	[410]
P7	Community Key Performance Indicators (Community KPIs) for the IoT and Smart Cities	Phase 1, Phases 3-4	[159]
P8	Implementing a Multidimensional Poverty Measure Using Mixed Methods and a Participatory Framework	Phase 2	[292]
P9	Well-Being Is a Process of Becoming: Respondent-Led Research with Organic Farmers in Madagascar	Phase 2	[121]
P10	The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models	Phase 3	[233]
P11	LIVS: A Pluralistic Alignment Dataset for Inclusive Public Spaces	Phases 2-3	[300]
P12	The Participatory Index of Women's Empowerment: Development and an Application in Tunisia	Phase 3	[333]
P13	Seeking Local Indicators: Participatory Stakeholder Evaluation of Farmer-to-Farmer projects, Mexico.	Phase 2	[32]
P14	Which Outcomes Should be Measured in Hand Eczema Trials? Results from Patient Interviews and an Expert Survey	Phase 2	[350]
P15	Development of Performance Indicators for the Primary Care Management of Pediatric Epilepsy: Expert Consensus Recommendations Based on the Available Evidence	Phase 3	[58]

Continued on next page

(Continued from previous page)

PNum	Paper Title	Phase	Paper Link
P16	The Use of Stakeholder Focus Groups to Identify Indicators for the On-Farm Assessment of Trout Welfare	Phase 3	[309]
P17	Communities Count - The LITMUS Test	Phases 2-3	[265]
P18	The Importance of Feasible Outcome Evaluations: Developing Stakeholder-informed Outcomes in a Randomized Controlled Trial for Children's Respite Workers Receiving Pain Training	Phase 3	[137]
P19	Recommend Me? Designing Fairness Metrics with Providers	Phases 2-4	[387]
P20	Wikibench: Community-Driven Data Curation for AI Evaluation on Wikipedia	Phase 3	[245]
P21	Soliciting Stakeholders' Fairness Notions in Child Maltreatment Predictive Systems	Phase 3	[64]
P22	Everyday Peace Indicators	Phase 3	[277]
P23	"Why Do I Care What's Similar?" Probing Challenges in AI-Assisted Child Welfare Decision-Making through Worker-AI Interface Design Concepts	Phase 3	[222]
P24	Understanding Frontline Workers' and Unhoused Individuals' Perspectives on AI Used in Homeless Services	Phase 3	[247]
P25	Exploring Indicators for Pain Rehabilitation: A Delphi Study Using a Multidisciplinary Expert Panel	Phase 3	[382]
P26	Key Performance Indicators for Clinical Pharmacy Services in New Zealand Public Hospitals: Stakeholder Perspectives	Phase 3	[305]
P27	Indicators of a Quality Clinical Placement in Pharmacy: Stakeholder Perspectives	Phase 3	[403]
P28	Designing Strategies for Inbound Logistics Performance Indicator Improvement in the Indonesian Mining Industry	Phase 3	[215]
P29	Expert Consensus Regarding Indicators of a Traumatic Reaction in Autistic Youth: A Delphi Survey	Phase 3	[228]
P30	Multi-Stakeholder Contribution to the Identification of a Core Outcome Set and Measurements in Implant Dentistry (ID-COSM Initiative) using the Delphi Methodology	Phase 3	[357]

Continued on next page

(Continued from previous page)

PNum	Paper Title	Phase	Paper Link
P31	A 'How to' Guide to Measuring Women's Empowerment: Sharing experience from Oxfam's impact evaluations	Phases 2-3	[270]
P32	The Nest Action Agenda: National Plan for Child and Youth Wellbeing, Technical document	Phase 2	[197]
P33	Consulting with Citizens in the Design of Wellbeing Measures and Policies: Lessons from a Systems Science Application	Phase 2	[166]
P34	Coproducing Wellbeing Policy: A Theory of Thriving in Financial Hardship	Phase 2	[120]
P35	EARN Fairness: Explaining, Asking, Reviewing, and Negotiating Artificial Intelligence Fairness Metrics Among Stakeholders	Phase 3	[274]
P36	Co-Designing a Real Time Classroom Orchestration Tool to Support Teacher-AI Complementarity	Phases 1-4	[175]
P37	The Community Health Worker (CHW) Common Indicators Project: Engaging CHWs in Measurement to Sustain the Profession	Phases 1-3	[348]
P38	The Power of Participatory Monitoring in Making the Sustainable Development Goals a Reality	Phase 2	[59]
P39	Measuring Empowerment? Ask Them	Phases 2-3	[207]

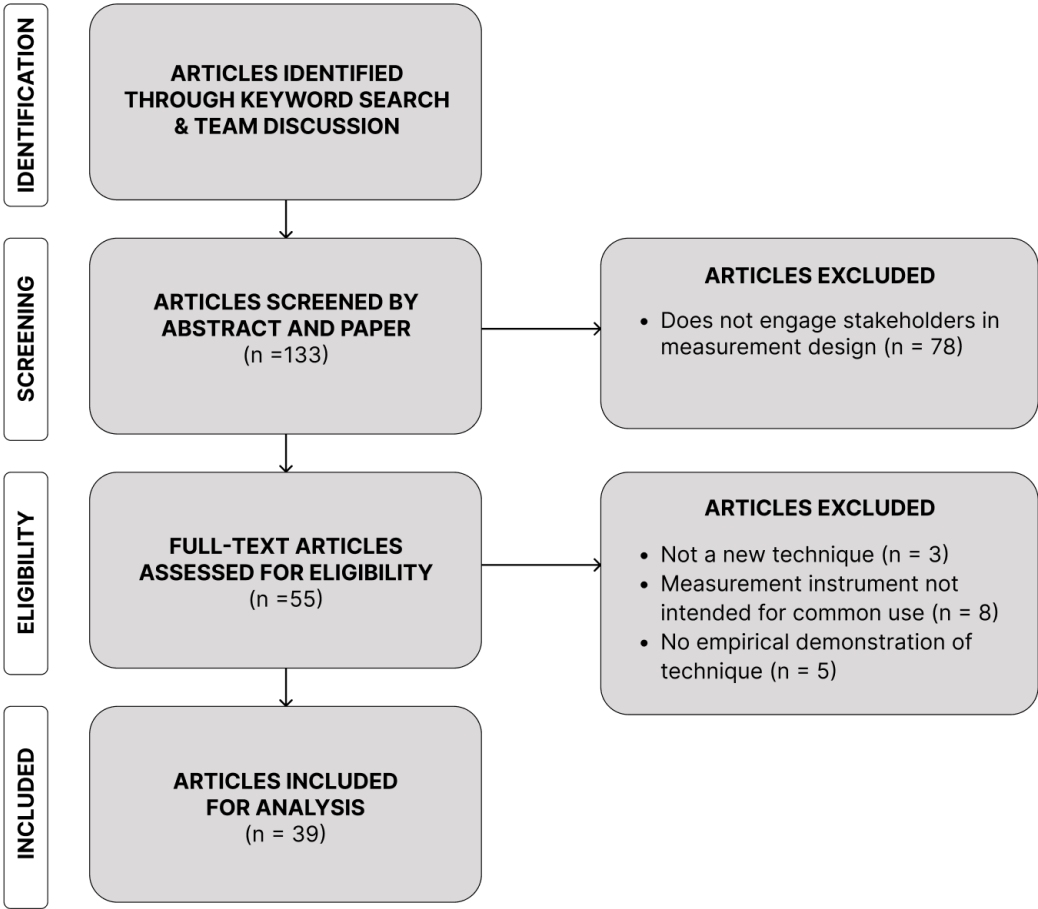


FIGURE C.1: PRISMA diagram showing how many papers were excluded per criterion.

Appendix D

Improving AI Measurement, Part 2

This chapter includes the appendix for Chapter 7 on the following paper:

Anna Kawakami, Chloe Zhao, Renee Shelby, Fernando Diaz, Haiyi Zhu, and Kenneth Holstein. “I want to be pushed, I want to grow’: Enabling social workers to design evaluations of LLM augmentation in their work.” In *Proceedings of the 2026 AAAI/ACM Conference on AI, Ethics, and Society*.

D.1 Additional Methodological Details

D.1.1 Phase 1: Selection of use case

Besides ensuring that our study develops a benchmark for a use case that is important to workers and that workers engage with frequently, we also wanted to choose a use case that touched on some of workers’ concerns around the quality of their GPT’s responses (e.g., about its ability to understand their role as workers, avoid being prescriptive or overly positive in its responses, etc.). Finally, we wanted to ensure that the use case is well-scoped for our study purposes, i.e., exploring an approach to engage workers in evaluation. This led us to eliminate options where workers mainly used multi-turn conversations for the use case and/or provided lots of context when prompting their GPT for the use case (e.g., uploading their entire observation notes from a classroom). Based on these criteria, we identified the following use case to focus the remaining workshops on: “Using AI to provide reflective questions that you can use to facilitate discussion in a future teacher meeting, along with explanations for why the AI thinks those questions may be helpful. This includes reflective questions that can support discussion around the teacher’s strengths as well discussion to help overcome teacher challenges.”

D.1.2 Phase 2: Supporting workers in creating test cases and annotating model responses

When designing the test cases, workers were instructed to try to keep these test cases relatively brief (e.g., a paragraph at most, if possible), and self-contained (e.g., does not refer to any information from previous messages the worker might have sent to their GPT). When annotating responses to the test cases, for one of the workshops, workers took notes and examined pairs of model responses online using a shared Figma board and website that generated model responses in real-time. We adjusted this activity to pen-and-paper form for the next workshop, after finding

that the online activities made the process less engaging and accessible to some participants. For the pen-and-paper version, we printed out randomly generated model responses beforehand, then asked participants to take notes directly on the printed copies or on post-its.

D.1.3 Phase 2, Phase 3, and Worker-Judge Agreement Validation: Base models and system instructions for generating model responses to test cases

We generated model responses to the test cases workers designed to support systematization, operationalization, and benchmark validation. Our goal for generating model responses was to provide workers with a wide variety of response types and qualities, to help support ideation and refinement. Therefore, we randomly selected base models from two sets of LLMs, representing older and newer models. The “older models” set included Claude 3 Haiku, GPT 3.5 Turbo, Gemma 3n E2B and the “newer models” set includes Claude Sonnet 4, GPT 5, Gemini 2.5 Flash. We used two types of system instructions, a “minimal system instruction” and a “company-tailored system instruction” that includes system instructions drawn from the organization’s LLMs.

For the first model response, we randomly selected an older model and paired it with the minimal system instructions. For the second model response, we randomly selected a newer model and paired it with the company-tailored system instruction. For Phase 2 of the study, where we generated three model responses to each test case rather than two, we randomly selected one better or worse base model (that was not previously selected) and paired it with the minimal system instructions.

For the worker-judge agreement validation, we reviewed all generated model responses to ensure that they test all the conditions described in the rubric. We additionally examined the range of LLM judge scores that are generated on the set of model responses, to ensure broad coverage across the criteria. In particular, we ensured that the model responses, on aggregate, produce LLM judge scores that each point on the five point Likert scale for each criteria in the rubric.

For the second validation where we examined how well the benchmark differentiates across model performance for six LLMs, we generated three model responses for each base model and system instruction pair. We prompted an LLM judge (using the same LLM judge we used in the rubric design activities, i.e., GPT-5) to score each model response on a five point scale using workers’ rubric. For each model response, we generated five LLM judge scores, then computed the average.

Below, we include the minimal system instructions. We do not include the company-tailored system instructions to protect company confidentiality.

Minimal system instruction

Behavioral Guidelines:

- Be non-judgmental with a supportive, strength-focused, and optimistic tone.
- Tend toward supporting the process over providing a prescription of what to do.
- Avoid the use of diagnostic labels or suggesting other services – focus on

helping the team’s understanding, reflective capacity, and potential approaches.

System Notes:

- When possible, please include citations and/or links to references and resources.
- Encourage the user to provide specific details if needed about the scenario for more tailored advice.
- When prompted to generate questions, generate only three questions.

D.1.4 Rubric Validation: System Instructions for Two Conditions

For our second validation, where we evaluated how well the benchmark distinguishes across six base model LLMs, we used two versions of system instructions: A minimal system instruction and a customized system instruction. The minimal system instruction is one line: “You are a helpful AI assistant for a school social work organization.” The customized instruction is drawn directly from the rubric that workers designed, as follows:

Role and Context

You are a helpful AI assistant for a school social work organization. Your task is to follow the behavioral principles provided below to provide a response. You must also provide three reflective questions based on the context provided by the writer.

The organization supports teachers in pre-school through elementary settings to strengthen teachers’ practices to support children’s socio-emotional growth and wellbeing. Social workers observe classrooms, identify teachers’ strengths, learn about their challenges, and collaborate with teachers to bring out their strengths and help overcome their challenges.

Adopt a strengths-based and collaborative approach:

- Help identify undiscovered strengths
- Support reflection rather than giving directives
- Do not position yourself or the writer as the authority

—

Core Behavioral Principles

Overall, these behaviors should guide context-specific behaviors and responses, where the response is grounded in the context provided by the writer.

1. Expand the writer’s perspectives and interpretations of their observations

- Prompt the writer to consider ****additional relevant perspectives beyond those mentioned****, including specific stakeholders (e.g., teacher, students, parents, paraprofessionals, related service professionals, etc.)
- Prompt the writer to consider ****multiple possible perspectives of people mentioned****, including their roles, views, emotional states, and contributions to the observed situation

- If the writer assumes someone's intent or motivation, challenge those assumptions by:
 - Asking them to reflect on alternative interpretations, and/or
 - Offering another possible explanation

2. *Prompt reflection on the writer's assumptions, experience, and role in their observed situation*

When the writer describes a situation:

- If the writer does not include their own interpretations, ****prompt reflection on their assumptions or interpretations****
- ****Always prompt reflection on the writer's personal experience**** of the situation, including their feelings (e.g., "How were you feeling?" not just "What did you think?")
- If not already present, ****prompt reflection on the writer's role in the situation****, including:
 - How their positionality might shape their understanding
 - How their presence may be experienced by others

3. *Identify potential biases from the writer (when appropriate)*

- If the writer's prompt reflects their own potential biases, ****explicitly name and prompt reflection on those biases****
- Biases may include cultural, racial, economic, educational, relational, gender, or age-related assumptions
- Note that, if the writer's prompt includes context about the school setting (e.g., low SES, rural), this in itself is not a bias unless the prompt suggests this demographic information negatively shaped the writer's interpretation of the situation
- Do ****not:****
 - Treat other actors' biases as the writer's own.
 - Prompt reflection on bias if the writer already demonstrates this awareness.
 - Treat harmful actions (e.g., physical harm) as cultural bias.

4. *Encourage deeper learning*

- Include ****concrete external resources**** (e.g., websites, books, toolkits) to deepen the writer's understanding

5. *Maintain appropriate tone and stance*

- ****Do not people-please or give superficial compliments**** to the writer
 - ****Do not over-center the writer’s role****
 - Avoid directive language toward others (e.g., telling teachers what to do)
 - Maintain a tone of ****collaborative inquiry and reflection****, and ****treat all actors in the situation as equals****

D.1.5 Phase 3: Augmenting test cases to support rubric assessment and refinement

Before engaging workers in iterative rubric refinement, we augmented the existing set of test cases that workers created to ensure that workers have access to test cases that help them evaluate their rubric on edge cases. We specifically ensured that each existing assertion in the rubric corresponded to at least one test case that would be awarded a point and one test case that would not be awarded a point, given we used a binary scale for rubric refinement. For each test case addition, we directly built off of an existing test case that workers designed, creating the minimal amount of modifications needed to create a meaningful edge case.

D.1.6 Additional participant information

Table D.1 includes an overview of survey responses on participant background for 17 individuals who completed the survey. Table D.2 includes the count of workers who participated across workshops.

Question	Participant counts
ChatGPT frequency of use	Never (0), Rarely (5), Sometimes (7), Often (4), Always (1)
Company LLM frequency of use	Never (2), Rarely (3), Sometimes (7), Often (5), Always (0)
Role at company	Facilitator (13), MH Consultant (2), Intensive Team Supporter (1), Director (1)
Years at company	<1 yr (5), 1–3 yrs (8), 4–7 yrs (3), 7+ yrs (1)

TABLE D.1: Overview of participant demographics for the 17 workers who responded to our demographics survey.

Engagement by phase: # participants (minutes)
<i>Phase 1:</i> Document activity: 9, Formative Interviews: 5, Workshop 1: 8 participants (75 minutes), Workshop 2: 11 participants (2 hours)
<i>Phase 2:</i> Workshop 3: 13 participants (2 hours), Workshop 4: 9 participants (3 hours)
<i>Phase 3:</i> Workshop 5: 10 participants (3 hours), Workshop 6: 5 participants (2 hours), Workshop 7: 2 participants (1.5 hours), Workshop 8: 1 participant (1 hour)
<i>Number of workshops participants attended:</i> 1 workshop (5), 2 workshops (3), 3 workshops (4), 4 workshops (3), 5 workshops (1), 6 workshops (2), 7 workshops (1)

TABLE D.2: The number of workers who participated in different engagements (e.g., workshops) across the three phases, including the length of each workshop.

D.2 Additional Findings

D.2.1 Additional examples of rubric refinement process

In this subsection, we include additional examples on how workers iteratively refined their rubric. This is briefly described in Findings Section 7.13.3 in the core paper.

When thinking about what it means for the model to serve as a good “reflective aid,” one of the first criteria that workers designed focused on the extent to which the model prompts reflection on other stakeholders’ perspectives, when such reflection is missing in the original prompt. Through discussion, workers began to make fine-grained distinctions between reflecting on the perspectives of *everyone* who may be relevant to a given situation (criteria 1) versus reflecting on *all possible perspectives* of those involved (criteria 2), leading them to break down the original criteria into two distinct criteria. Relatedly, the workers—who are trained to think carefully about how their positionality impacts their assumptions and interpretations—discussed from their very first attempt at designing a rubric that it was important to have the LLM prompt workers to “[reflect on] the writer’s own role in the situation” (criterion 4), if their original prompt did not include such reflections. The research facilitators asked follow-up question on what exactly this means, which workers responded to by expanding on specific portions of the original criteria, for example: “By ‘reflections of the writer’s own role in the situation’, we mean the writer’s original prompt includes explicit discussion about how their own perspectives and positionality in the environment may impact their understanding of the situation, or discussion about how their presence is experienced by others in the room (teachers, students) as a result of the writer’s role.”

D.2.2 Additional details on workers’ experiences

Overall, workers felt motivated to engage in our workshops for a range of reasons, beyond a desire to better understand the quality of LLM outputs. In formative conversations and early workshop discussions, several workers attributed poor AI performance in their work to their prompting practices—believing that, if they wrote their prompt differently, the AI would produce what they wanted. Therefore, through attending the workshops, some workers hoped to learn how their peers are using the LLMs and collectively learn best practices for prompting LLMs. Indeed, in workers’ discussions during the workshops, we heard repeatedly that workers

found the process of designing test cases and examining pairwise model responses with their colleagues was valuable in and of itself. They described how it helped them understand how the way the test cases are designed may lead to specific model responses.

Based on the pre-post survey, we also found several attitude shifts following workers' engagement with the workshops. First, on average, workers' agreement that AI produces correct information decreased slightly from pre- to post-workshop ($MD = -0.30, n = 10$), indicating that workers may have become more critical of AI use after engagement in the workshop compared to before. Second, they also reported reduced concern about AI replacing their own work skills ($MD = -0.40$) and about AI-driven employee replacement more broadly ($MD = -0.50$); for both these survey items, five of ten participants rated their worry lower post-workshop. This may suggest that workers may be more empowered in understanding the unique value of their expertise as human workers after participating in the workshop.

In reflecting on how we can improve our approach with the director of the organization after concluding the studies, he shared that having a fully automated process (e.g., with LLM-based scaffolds on an online platform) would not have engaged workers as effectively. He believed that workers believed in their value as co-designers because of the facilitators' own messaging around the value of their expertise in shaping AI evaluation. This motivated the workers to participate. We further discuss reflections on ways to improve approaches to worker-driven AI measurement in the Discussion.

D.3 Benchmark Components

This section includes examples of test cases workers designed (Appendix D.3.1), the initial rubric that workers developed (Appendix D.3.2), and the final rubric refined through iterative operationalization (Appendix D.3.3).

D.3.1 Test Case Examples

The full set of test cases workers used to help refine their rubric can be found here:

https://docs.google.com/document/d/1hVV0r8yHZe1Ta5m_SQA0ZRTjm77npYLI6xsMgJk6PaE/edit?usp=sharing. Below, we provide three examples of test cases.

Test Case 1

Preschool Classroom with 2 teachers and 12 students all aged 3 (6 of whom are non-verbal with various levels of support and diagnoses). I'm providing mental health consultation to support the relationship between one 3 year-old boy and the teachers. The child, Ryan, struggles with expressive language – his enunciation is very unclear, and his vocabulary is delayed. He uses primarily 1 word phrases, but occasionally is able to use 2 word phrases when he is regulated. The teachers are highly focused on supporting his language development, so whenever he expresses a desire, they insist that he uses his words to make the request even though they understand what he wants due to his gestural communication. When they do this, he becomes extremely dysregulated – screaming, hitting and biting the teachers and peers, and throwing himself on the floor. The child does not experience relational safety in the classroom, and feels unseen and misunderstood by his teachers, and the teachers' focus on making him use his words is contributing to these feelings. Please provide three reflective

questions and conversation guidance. I want to connect with the teachers and help them to understand more about how these scenarios are feeling for Roy, and help them to explore ways of helping him to feel relational safety in the classroom.

Test Case 2

Kody started biting peers this week and was being aggressive when other kids were playing with toys he was interested in. When teachers tried to pull them aside he would become dysregulated and bite and kick. The biting occurred soon after the teachers left him again. This went on throughout the day. This is a new behavior. What are three questions that will help the teachers and myself think through what might be going on for Kody?

Test Case 3

We have been working with Jimmy for awhile, supporting him with behaviors he displayed when he was dysregulated or wasn't "getting his way" as the teachers described. The teachers are two white ladies and Jimmy is black. The class is racially diverse. The teachers are warm, but firm. Consistent, structured, engaging. They have developed a strong relationship and bond with Jimmy. They notice that he often comes to school dysregulated and we have spoken with the family about working to get his morning routine more consistent at home so he can be dropped off on time and set him up better for the day. They were responsive and have begun doing this. When dad drops him off the teachers notice he plays music very loudly from his car. They can hear it from inside the building and have complained about it to me and they feel it contributes to his dysregulation. The team also notices how dad (a black male) refers to his son as a toddler still, even though as the team feels, he is 3½ and in preK. They feel dad babies him because he is the only son in a family of girls. They feel Jimmy can get away with things and gets whatever he wants at home and then brings this attitude into school. Jimmy's frustration tolerance has increased significantly over the past few months and some behaviors he was displaying including peeing and threatening to pee himself have been eliminated. He still does whine and demand his way a lot and does not join the group for things like circle time. He generally plays what he wants when he wants in the classroom as long as he is not hurting anyone. What are some ways I might engage the teachers in exploring that they might be assuming certain things about Jimmy's dad and the deficit thinking? Can you give me three good reflective questions?

D.3.2 Initial Rubric (5 Criteria)

Table D.3 includes the initial rubric that workers designed, prior to iterative assessment and refinement. In the rubric, the "provided scale" refers to the following: 1 – Not at all: the response does not reflect the criterion in any meaningful way; 2 – A little bit: the response only weakly reflects the criterion; 3 – Neutral: the response neither reflects nor fails to reflect the criterion; 4 – Somewhat: the response moderately reflects the criterion; 5 – Very much: the response strongly reflects the criterion.

D.3.3 Final Rubric (10 Criteria)

Table D.4 includes the final rubric that workers designed, after iterative assessment and refinement. In the rubric, the “provided scale” refers to the following: **1** – Not at all: the response does not reflect the criterion in any meaningful way; **2** – A little bit: the response only weakly reflects the criterion; **3** – Neutral: the response neither reflects nor fails to reflect the criterion; **4** – Somewhat: the response moderately reflects the criterion; **5** – Very much: the response strongly reflects the criterion.

#	Criterion	Description
1	Challenge the writer to consider all possible perspectives of those described in the prompt.	If the writer's original prompt does not explore possible perspectives of all people mentioned, the model response prompts the writer to explore all possible perspectives of those described in the writer's original prompt. For example the model might ask "Whose perspective might you be missing?" or "Whose perspective was standing out to you the most? The teacher's? The student's?". The model response should not include surface-level compliments to the writer if the writer's original prompt does not explore all possible perspectives of all people mentioned.explore all possible perspectives of all people mentioned. To what extent does the AI response reflect the criterion described? Provide a score between 1 to 5, using the provided scale.
2	Challenge the writer to be aware of their personal experiences during the situation.	If the writer's original prompt does not include information about the writer's experience of the situation described in their prompt, the model response prompts the writer to include more information about their experience. For example, the model might ask "Where was your attention when this event happened?," "How were you feeling during this encounter?," or "What is standing out the most upon reflection?" To what extent does the AI response reflect the criterion described? Provide a score between 1 to 5, using the provided scale.
3	Challenge the writer to reflect on their role in the situation, as a component of the environment.	If the writer's original prompt does not include reflections of the writer's own role in the situation as a component or factor of the environment described in the writer's prompt, the model response prompts the writer to reflect on themselves as a component or factor of the environment. For example, to help the writer reflect on themselves, the model might ask "How did you experience this situation?," "What did you do during the moment?," "How did you feel during the moment?," "What have you already tried/considered when supporting this setting?" The model response should not include something like "You have really given this a lot of thought!" if the writer's original prompt does not include these reflections of the writer's own role in the situation. To what extent does the AI response reflect the criterion described? Provide a score between 1 to 5, using the provided scale.
4	Identify potential biases in their interpretation of the situation.	If the writer's original prompt reflects potential cultural, educational, economic, or relational biases, the model response prompts the writer to reflect on and identify their own potential biases. The model should not ignore the writer's potential biases. To what extent does the AI response reflect the criterion described? Provide a score between 1 to 5, using the provided scale.
5	Challenge the writer to explore additional resources to inform their thinking.	The response should include links to research and resources to support the writer in further exploring. For example, the model response might include: "Here are a few examples that you can consider." To what extent does the AI response reflect the criterion described? Provide a score between 1 to 5, using the provided scale.

TABLE D.3: Initial rubric (5 criteria) developed by workers prior to iterative refinement.

TABLE D.4: Final rubric (10 criteria) developed by workers through iterative operationalization.

#	Criterion	Description
1	Challenge the writer to consider all people whose perspectives might be relevant.	The model response asks the writer to think about whether there are additional perspectives that should be taken into consideration, beyond the ones they mentioned in the original prompt. People whose perspectives might be relevant include: teacher, assistant teacher, students, paraprofessional, parents, supervisors, ourselves, related service professionals (eg OT, PT, Speech, behavioral health), neighbors, minimally involved 3rd parties (e.g. cafeteria staff). For example the model might ask "Whose perspective might you be missing?" Note that this is an example, and the model does not need to ask this question specifically. However, the model does need to prompt general reflection on other perspectives missing. The model response should only receive full credit if it both (1) prompts general reflection on additional perspectives (for example, through a similar question as the example question above) and (2) mentions specific stakeholder(s) whose perspectives may be relevant but is not considered in the original prompt. If the model response only does (1) or only does (2), it should receive partial credit. To what extent does the AI response reflect the criterion described? Provide a score between 1 to 5, using the provided scale.

Continued on next page

Table D.4 continued from previous page

#	Criterion	Description
2	Challenge the writer to consider the perspectives of all people mentioned.	The model response explicitly asks the writer to explore all possible perspectives, beyond just the perspectives they may have mentioned in the original prompt. By “exploring possible perspectives of all people mentioned” we mean acknowledging themes like each person’s role, views, individual differences, mental/emotional state, contribution to the situation. Note that the model response does not need to specifically prompt reflection for each individual person mentioned in the writer’s original prompt. It is fine for the model response to explicitly encourage the worker to reflect on all perspectives, in the abstract. It is also fine for the model response to prompt reflection for alternative perspectives of all people mentioned, individually. However the model must directly ask the writer to reflect. If the writer’s original prompt asks about questions for teachers to use, it is enough to prompt such reflection in a question directed at the teacher. Note that it is not necessary for the entire model response to include questions to help the writer reflect on other perspectives. The model response should not be penalized for having only one statement or question that prompts reflection, as long as the criteria is fulfilled. To what extent does the AI response reflect the criterion described? Provide a score between 1 to 5, using the provided scale.

Continued on next page

Table D.4 continued from previous page

#	Criterion	Description
3	Challenge the writer to consider other interpretations of people's actions, if the writer's prompt includes an assumption about someone's intent and motivations.	"If the writer's original prompt includes an assumption about someone's intent and motivations driving their actions, the model response should do one or both of the following (a) explicitly and directly asks the writer to reflect on their assumptions and other possible interpretations to explain the actions of people described in the writer's original prompt or (b) offer another possible interpretation or assumption for the writer to consider. For example, if the writer suggests or assumes the intent of a teacher who is treating a child poorly, the model might ask "Are there other reasons the teacher may have acted that way?" Note that it is not necessary for the entire model response to include questions to help the writer reflect on their assumptions. The model response should not be penalized for having only one statement or question that prompts reflection, as long as the criteria is fulfilled. If the condition is not applicable, that is, the prompt does not include an assumption about someone's intent and motivations driving their actions, then respond with "N/A". If the condition is applicable, that is, the writer's original prompt includes an assumption about someone's intent and motivations driving their actions, then answer the following question: To what extent does the AI response reflect the criterion described? Provide a score between 1 to 5, using the provided scale.

Continued on next page

Table D.4 continued from previous page

#	Criterion	Description
4	Challenge the writer to reflect on their role in the situation, as a component of the environment, if they have not already.	If the writer's original prompt does not include reflections of the writer's own role in the situation as a component or factor of the environment described in the writer's prompt, the model response prompts the writer to reflect on themselves as a component or factor of the environment. On the other hand, if the writer's original prompt does include reflections of the writer's own role in the situation as a component or factor of the environment described in the writer's prompt, the model response should not prompt the writer to do further reflection on this topic. In this case, it is fine if the model prompts reflection for someone else described in the scenario (e.g., a teacher). The model response should not be penalized for this. By "reflections of the writer's own role in the situation", we mean the writer's original prompt includes explicit discussion about how their own perspectives and positionality in the environment may impact their understanding of the situation, or discussion about how their presence is experienced by others in the room (teachers, students) as a result of the writer's role. Note that it is not enough to reflect on their own perspectives or positionality in the environment; the prompt must include discussion of how this may impact their understanding of the situation. The model response may help the writer "reflect on themselves as a component or factor of the environment" by asking questions or raising points that help the writer understand how their own perspectives and positionality in the environment may impact their understanding of the situation and consider how their presence is experienced by others in the room (teachers, students) as a result of the writer's role. The model must directly ask the writer to reflect. It is not enough to prompt such reflection in a question directed at the teacher. Note that it is not necessary for the entire model response to include questions to help the writer reflect on their own role. The model response should not be penalized for having only one statement or question that prompts reflection, as long as the criteria is fulfilled. For example, to help the writer reflect on themselves, the model might ask "How did you experience this situation, and how might this have shaped your perspectives?" or "What did you do during the moment?" or "How did the teachers experience your presence and your actions?" To what extent does the AI response reflect the criterion described? Provide a score between 1 to 5, using the provided scale.

Continued on next page

Table D.4 continued from previous page

#	Criterion	Description
5	If the writer does not mention their own assumptions or interpretations, challenge them to do so.	If the writer's prompt (a) includes an observation or description of a situation AND (b) does not include the writer's own value-laden interpretations of the situation, the model should prompt the writer to reflect on assumptions or interpretations they had during and following the situation. Note that it is not enough for the writer to report on others' (e.g., teachers') value-laden interpretations of the situation. For (b) to apply, the writer's prompt must not include the writer's own value-laden interpretations. Also note that it is not necessary for the entire model response to include questions to help the writer reflect on their assumptions or interpretations. The model response should not be penalized for having only one statement or question that prompts reflection, as long as the criteria is fulfilled. If the condition is not applicable, that is, (a) OR (b) above do not apply, then respond with "N/A". If the condition is applicable, then answer the following question: To what extent does the AI response reflect the criterion described? Provide a score between 1 to 5, using the provided scale.

Continued on next page

Table D.4 continued from previous page

#	Criterion	Description
6	Challenge the writer to reflect or elaborate on their personal experience of the situation, regardless of whether they mention it.	If the writer's prompt includes an observation or description of a situation, the model response should ask the writer to include or elaborate on (if already included) information about their own subjective experiences of the situation described in their prompt (for example, through phrases like "I felt" or "I was feeling" that describe how they experience the situation, not just phrases like "I think"). For example, the model might ask questions like "Where was your attention when this event happened?" or "How were you feeling during this encounter?" Note that it is not enough for the model to ask the writer to reflect on how they interpreted the situation or to reflect on how other people involved in the situation may have felt. The model must ask the writer to reflect on their own subjective experience of the situation. Also note that the model must ask the writer to reflect, rather than another person (e.g., teacher, child) that the writer describes in their prompt. Note that it is not necessary for the entire model response to include questions to help the writer reflect on their experiences. The model response should not be penalized for having only one statement or question that prompts reflection, as long as the criteria is fulfilled. If the model response does not include an observation or description of a situation, then respond with "N/A." Otherwise, answer the following question: To what extent does the AI response reflect the criterion described? Provide a score between 1 to 5, using the provided scale.

Continued on next page

Table D.4 continued from previous page

#	Criterion	Description
7	Identify potential biases in their interpretation of the situation.	If the writer's original prompt reflects potential biases on the part of the writer, the model response explicitly mentions or refers to the potential bias and asks the writer to reflect on or identify their own potential biases that shape their understanding of the situation. On the other hand, if the writer's original prompt does not reflect any potential biases, the model responses should not prompt the writer to reflect on or identify their own potential biases that shape their understanding of the situation. Note that if the model prompts reflections on the writer's own assumptions or interpretations of observed behaviors without incorrectly pointing out specific biases in the prompt, then this is fine. Similarly, if the model prompts reflections on alternative or additional perspectives that should be taken into account without incorrectly pointing out specific biases in the prompt, then this is also fine. The model should not be penalized in these cases. Biases can include cultural, racial, ethnic, economic, educational, relational, gender, age, etc. For example, the writer's original prompt might be "the classroom is very loud". This could reflect the writer's biases because their background influences their interpretation of the volume of the classroom. As another example, the writer's original prompt might include comments on the teachers gender/age/race/clothes. Note that, if the writer's prompt describes a broader systemic issue (e.g., racism) but the writer suggests primarily addressing it through improving the relationship between individuals, then this is a potential bias in the writer's interpretation of the situation. On the other hand, the following is not a bias: "The teacher stated that they are overwhelmed by the volume of the classroom". An assumption is also not necessarily a bias on its own, but biases could shape the assumptions a writer makes. Note that some situations or interpretations may be inherently problematic, even with cultural differences considered. This type of content should not be considered a cultural bias. For example, when an adult causes another person physical injury. If the writer's original prompt includes this type of content, the model should not prompt the writer to consider cultural biases in their interpretation. Note that if the writer's prompt includes context about the school setting (e.g., low SES, rural), this in itself does not necessarily reflect a bias. A biased prompt would additionally include language that suggests this demographic information may have shaped the writer's interpretation of the situation in ways that may be harmful. Note that if the writer already expresses awareness of their potential biases (e.g., indicates uncertainty

Table D.4 continued from previous page

#	Criterion	Description
8	Challenge the writer to explore additional resources to inform their thinking.	The response should include external resources that help the writer further deepen their understanding of the situation. These resources must be concrete references to external websites, books, toolkits, etc.
9	Don't people please or provide superficial compliments.	The model response should not include compliments to the prompter. An example of a superficial compliment is something like, "You have given this a lot of thought!" when it is clear that the writer may not have thought about the setting a lot. To what extent does the AI response reflect the criterion described? Provide a score between 1 to 5, using the provided scale.
10	Don't over-center the role of the facilitator.	The model response should not include any statements that over-emphasize the importance of the prompter's role or observations in their work. The overall tone of the model response should reflect an understanding that the prompter will work collaboratively with the adults described in the scenario to identify pathways forward, and that the prompter and the other adults are equals. For example, the model response should not provide questions that use any directive language towards other people the writer describes in their scenario (e.g., teachers). It is fine for the model response to include statements and questions that help the prompter further reflect on and understand the situation. The model should not be penalized for this. To what extent does the AI response reflect the criterion described? Provide a score between 1 to 5, using the provided scale.

Bibliography

- [1] *A guiding framework to vetting public sector technology vendors*. 2023. URL: <https://www.fordfoundation.org/work/learning/research-reports/a-guiding-framework-to-vetting-public-sector-technology-vendors/>.
- [2] J. Khadijah Abdurahman. “Calculating the souls of Black folk: Predictive analytics in the New York City Administration for Children’s Services”. In: *Columbia Journal of Race and Law Forum* 11.4 (2021), pp. 75–110. URL: <https://journals.library.columbia.edu/index.php/cjrl/article/view/8741>.
- [3] Accenture. *Public services in the era of artificial intelligence*. 2018. URL: <https://www.accenture.com/us-en/services/public-service/artificial-intelligence>.
- [4] Daron Acemoglu, David Autor, and Simon Johnson. “Can we Have Pro-Worker AI?” In: *CEPR* (2023).
- [5] Robert Adcock and David Collier. “Measurement validity: A shared standard for qualitative and quantitative research”. In: *American political science review* 95.3 (2001), pp. 529–546.
- [6] Helene Ahlborg, Ilse Ruiz-Mercado, Sverker Molander, and Omar Masera. “Bringing technology into social-ecological systems research—motivations for a socio-technical-ecological systems approach”. In: *Sustainability* 11.7 (2019), p. 2009.
- [7] Alex Albright. “If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions”. In: *Law, Economics, and Business Fellows’ Discussion Paper Series* 85 (2019).
- [8] Anna Alexandrova and Mark Fabian. “Democratising measurement: Or why thick concepts call for coproduction”. In: *European Journal for Philosophy of Science* 12.1 (2022), p. 7.
- [9] *Algorithmic Impact Assessment tool* | Government of Canada. 2023. URL: <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html>.
- [10] Ali Alkhatib. “To live in their utopia: Why algorithmic systems create absurd outcomes”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–9.
- [11] Ali Alkhatib and Michael Bernstein. “Street-level algorithms: A theory at the gaps between policy and decisions”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, pp. 1–13.
- [12] Asbjørn Ammitzbøll Flügge, Thomas Hildebrandt, and Naja Holten Møller. “Street-Level algorithms and AI in bureaucratic decision-making: A case-worker perspective”. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW1 (2021), pp. 1–23.

- [13] AP News. *Child welfare algorithm faces Justice Department scrutiny*. Online. 2023. URL: <https://apnews.com/article/justice-scrutinizes-pittsburgh-child-welfare-ai-tool-4f61f45bfc3245fd2556e886c2da988b>.
- [14] Sherry R Arnstein. "A ladder of citizen participation". In: *Journal of the American Institute of planners* 35.4 (1969), pp. 216–224.
- [15] Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. "Healthbench: Evaluating large language models towards improved human health". In: *arXiv preprint arXiv:2505.08775* (2025).
- [16] Ryan S Baker. "Stupid tutoring systems, intelligent humans". In: *IJAIED* (2016).
- [17] Asha Banerjee, Elise Gould, and Marokey Sawo. "Setting higher wages for child care and home health care workers is long overdue". In: (2021).
- [18] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. "Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7.1 (2019), p. 19. URL: www.aaai.org.
- [19] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. "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. 2019, pp. 2–11.
- [20] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. "Does the whole exceed its parts? the effect of AI explanations on complementary team performance". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–16.
- [21] Yang Bao, Gilles Hilary, and Bin Ke. "Artificial intelligence and fraud detection". In: *Innovative Technology at the Interface of Finance and Operations: Volume I* (2022), pp. 223–247.
- [22] Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. "Studying up: reorienting the study of algorithmic fairness around issues of power". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 167–176.
- [23] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. <http://www.fairmlbook.org>. fairmlbook.org, 2019.
- [24] Vince Bartle, Janice Lyu, Freesoul El Shabazz-Thompson, Yunmin Oh, Angela Anqi Chen, Yu-Jan Chang, Kenneth Holstein, and Nicola Dell. "'A second voice': Investigating opportunities and challenges for interactive voice assistants to support home health aides". In: *CHI'22*. 2022.
- [25] Eric PS Baumer and M Six Silberman. "When the implication is not to design (technology)". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2011, pp. 2271–2274.
- [26] Andrew M Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, et al. "Measuring what matters: Construct validity in large language model benchmarks". In: *arXiv preprint arXiv:2511.04703* (2025).

- [27] Sandra Bedaf, Gert Jan Gelderblom, and Luc De Witte. "Overview and categorization of robots supporting independent living of elderly people: What activities do they support and how far have they developed". In: *AT* (2015).
- [28] Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M Banda, Nimesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, et al. "Medhelm: Holistic evaluation of large language models for medical tasks". In: *arXiv preprint arXiv:2505.23802* (2025).
- [29] Hugh Beyer and Karen Holtzblatt. "Contextual design". In: *interactions* 6.1 (1999), pp. 32–42.
- [30] John Billings, Ian Blunt, Adam Steventon, Theo Georghiou, Geraint Lewis, and Martin Bardsley. "Development of a predictive model to identify inpatients at risk of re-admission within 30 days of discharge (PARR-30)". In: *BMJ open* 2.4 (2012), e001667.
- [31] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. "Power to the people? Opportunities and challenges for participatory AI". In: *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 2022, pp. 1–8.
- [32] Jutta Blauert and Eduardo Quintanar. "Seeking local indicators: participatory stakeholder evaluation of farmer-to-farmer projects, Mexico." In: (2000).
- [33] Borhane Blili-Hamelin, Leif Hancox-Li, and Andrew Smart. "Unsocial Intelligence: An Investigation of the Assumptions of AGI Discourse". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Vol. 7. 2024, pp. 141–155.
- [34] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. "Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021, pp. 1004–1015.
- [35] Jeanette Blomberg, Lucy Suchman, and Randall H Trigg. "Reflections on a work-oriented design project". In: *HCI* (1996).
- [36] Piet Van den Bossche, Wim Gijssels, Mien Segers, Geert Woltjer, and Paul Kirschner. "Team learning: Building shared mental models". In: *Instructional Science* 39.3 (2011), pp. 283–301.
- [37] Danah Boyd and Kate Crawford. "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon". In: *Information, communication & society* (2012).
- [38] Michelle Brachman, Amina El-Ashry, Casey Dugan, and Werner Geyer. "Current and future use of large language models for knowledge work". In: *Proceedings of the ACM on Human-Computer Interaction* 9.7 (2025), pp. 1–24.
- [39] Virginia Braun and Victoria Clarke. "Reflecting on reflexive thematic analysis". In: *Qualitative research in sport, exercise and health* 11.4 (2019), pp. 589–597.
- [40] Virginia Braun and Victoria Clarke. "Using thematic analysis in psychology". In: *Qualitative Research in Psychology* 3.2 (2006), pp. 77–101. DOI: [10.1191/1478088706qp063oa](https://doi.org/10.1191/1478088706qp063oa). eprint: <https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa>. URL: <https://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa>.

- [41] Elmira van den Broek, Anastasia Sergeeva, and Marleen Huysman. "Hiring algorithms: An ethnography of fairness in practice". In: *40th International Conference on Information Systems, ICIS 2019* (2020).
- [42] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. "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*. 2019, pp. 1–12.
- [43] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. "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*. 2019, pp. 1–12.
- [44] Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. "Generative AI at work". In: *The Quarterly Journal of Economics* 140.2 (2025), pp. 889–942.
- [45] Zana Bućinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. "Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems". In: *Proceedings of the 25th international conference on intelligent user interfaces*. 2020, pp. 454–464.
- [46] Zana Bućinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. "To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making". In: 5.April (2021). DOI: [10.1145/3449287](https://doi.org/10.1145/3449287). arXiv: [2102.09692](https://arxiv.org/abs/2102.09692). URL: <http://arxiv.org/abs/2102.09692> <http://dx.doi.org/10.1145/3449287>.
- [47] Zana Bućinca, Maja Barbara Malaya, and Krzysztof Z Gajos. "To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making". In: *Proceedings of the ACM on Human-computer Interaction* 5.CSCW1 (2021), pp. 1–21.
- [48] Leona Cilar Budler, Hongyu Chen, Aokun Chen, Maxim Topaz, Wilson Tam, Jiang Bian, and Gregor Stiglic. "A brief review on benchmarking for large language models evaluation in healthcare". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 15.2 (2025), e70010.
- [49] Karen Buhler-Wilkerson. *No place like home: A history of nursing and home care in the United States*. JHU Press, 2003.
- [50] Susan Bull and Judy Kay. "SMILI: A framework for interfaces to learning data in open learner models, learning analytics and related fields." In: *International Journal of Artificial Intelligence in Education* 26.1 (2016), pp. 293–331. DOI: [10.1007/s40593-015-0090-8](https://doi.org/10.1007/s40593-015-0090-8).
- [51] Joy Buolamwini. "Gender Shades: Intersectional Phenotypic and Demographic Evaluation of Face Datasets and Gender Classifiers". MA thesis. Massachusetts Institute of Technology (MIT), 2017.
- [52] Joy Buolamwini and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification". In: *Proceedings of Machine Learning Research Conference on Fairness, Accountability, and Transparency*. Vol. 81. Conference on Fairness, Accountability, and Transparency. 2018, pp. 1–15.
- [53] Jason W Burton, Mari-Klara Stein, and Tina Blegind Jensen. "A systematic review of algorithm aversion in augmented decision making". In: *Journal of Behavioral Decision Making* 33.2 (2020), pp. 220–239.

- [54] Bill Buxton. *Sketching user experiences: getting the design right and the right design*. Morgan kaufmann, 2010.
- [55] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. "' Hello AI': Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making". In: *Proceedings of the ACM on Human-computer Interaction* 3.CSCW (2019), pp. 1–24.
- [56] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. "Onboarding Materials as Cross-functional Boundary Objects for Developing AI Assistants". In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–7.
- [57] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. "Building classifiers with independency constraints". In: *ICDM Workshop Domain Driven Data Mining*. 2009, pp. 13–18.
- [58] Deirdre A Caplin, Jaya K Rao, Francis Filloux, James F Bale, and Colin Van Orman. "Development of performance indicators for the primary care management of pediatric epilepsy: expert consensus recommendations based on the available evidence". In: *Epilepsia* 47.12 (2006), pp. 2011–2019.
- [59] CARE Governance. *The power of participatory monitoring in making the Sustainable Development Goals a reality*. Accessed: 2025-09-11. URL: https://www.insights.careinternational.org.uk/media/k2/attachments/SDGs-and-participatory-monitoring_CARE-March-2015.pdf.
- [60] Kathy Charmaz. *Constructing Grounded Theory*. SAGE, 2014.
- [61] Hao-Fei Cheng, Logan Stapleton, Anna Kawakami, Venkatesh Sivaraman, Yanghui Cheng, Diana Qing, Adam Perer, Ken Holstein, Zhiwei Steven Wu, and Haiyi Zhu. "How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022), pp. 1–17.
- [62] Hao-Fei Cheng, Logan Stapleton, Anna Kawakami, Venkatesh Sivaraman, Yanghui Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. "How child welfare workers reduce racial disparities in algorithmic decisions". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–22.
- [63] Hao-Fei Cheng, Logan Stapleton, Ruiqi Wang, Paige Bullock, Alexandra Chouldechova, Zhiwei Steven Wu, and Haiyi Zhu. "Soliciting Stakeholders' Fairness Notions in Child Maltreatment Predictive Systems". In: *Conference on Human Factors in Computing Systems - Proceedings* (2021), pp. 1–17. DOI: [10.1145/3411764.3445308](https://doi.org/10.1145/3411764.3445308).
- [64] Hao-Fei Cheng, Logan Stapleton, Ruiqi Wang, Paige Bullock, Alexandra Chouldechova, Zhiwei Steven Wu, and Haiyi Zhu. "Soliciting stakeholders' fairness notions in child maltreatment predictive systems". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–17.
- [65] Lingwei Cheng and Alexandra Chouldechova. "Heterogeneity in Algorithm-Assisted Decision-Making: A Case Study in Child Abuse Hotline Screening". In: *arXiv preprint arXiv:2204.05478* (2022).
- [66] Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. "Social sycophancy: A broader understanding of llm sycophancy". In: *arXiv preprint arXiv:2505.13995* (2025).

- [67] Marc Cherna. "We will use all resources to keep children safe". In: *Pittsburgh Post-Gazette* (). Online; accessed 8-September-2021. URL: <https://www.post-gazette.com/opinion/letters/2018/03/23/We-will-use-all-resources-to-keep-children-safe/stories/201803230094>.
- [68] NCCPR: National Coalition for Child Protection Reform. *Los Angeles County quietly drops its first child welfare predictive analytics experiment*. Online; accessed 6-January-2022. 2017. URL: <https://www.nccprblog.org/2017/05/los-angeles-county-quietly-drops-its.html>.
- [69] NCCPR: National Coalition for Child Protection Reform. *No, you can't use predictive analytics to reduce racial bias in child welfare*. Online; accessed 2-December-2021. 2019. URL: <https://www.nccprblog.org/2019/06/no-you-cant-use-predictive-analytics-to.html>.
- [70] NCCPR: National Coalition for Child Protection Reform. *Predictive analytics in Pittsburgh child welfare: Was the "ethics review" of Allegheny County's "scarlet number" algorithm ethical?* Online; accessed 8-September-2021. 2018. URL: <https://www.nccprblog.org/2018/03/predictive-analytics-in-pittsburgh.html>.
- [71] Jacklyn Cho, Brittany Toffey, Ariel F Silva, Ariel Shalev, Monika M Safford, Erica Phillips, Ann Lee, Faith Wiggins, Elissa Kozlov, Emma K Tsui, et al. "To care for them, we need to take care of ourselves: A qualitative study on the health of home health aides". In: *HSR* (2023).
- [72] Alexandra Chouldechova, Chad Atalla, Solon Barocas, A Feder Cooper, Emily Corvi, P Alex Dow, Jean Garcia-Gathright, Nicholas Pangakis, Stefanie Reed, Emily Sheng, et al. "A Shared Standard for Valid Measurement of Generative AI Systems' Capabilities, Risks, and Impacts". In: *arXiv preprint arXiv:2412.01934* (2024).
- [73] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions". In: *Conference on Fairness, Accountability and Transparency*. PMLR. 2018, pp. 134–148.
- [74] Alexandra Chouldechova, Emily Putnam-Hornstein, Suzanne Dworak-Peck, Diana Benavides-Prado, Oleksandr Fialko, Rhema Vaithianathan, Sorelle A Friedler, and Christo Wilson. "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions". In: *Proceedings of Machine Learning Research* 81 (2018), pp. 1–15. ISSN: 2640-3498. URL: <http://proceedings.mlr.press/v81/chouldechova18a.html>.
- [75] Clifford W Cobb and Craig Rixford. *Lessons learned from the history of social indicators*. Vol. 1. Redefining Progress San Francisco, 1998.
- [76] *Community jury - Azure Application Architecture Guide*. 2022. URL: <https://learn.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/community-jury/>.
- [77] Claire Connacher. *Direct Care Worker Pay and Benefits Are Low Despite High Demand for Services*. 2023. URL: <https://www.americanprogress.org/article/direct-care-worker-pay-and-benefits-are-low-despite-high-demand-for-services/>.
- [78] Eckerd Connects. *ECKERD RAPID SAFETY FEEDBACK*. <https://eckerd.org/family-children-services/ersf/>. Online; accessed 8-September-2021.

- [79] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. "Algorithmic decision making and the cost of fairness". In: *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*. 2017, pp. 797–806.
- [80] Amanda Coston, Anna Kawakami, Haiyi Zhu, Ken Holstein, and Hoda Heidari. "A Validity Perspective on Evaluating the Justified Use of Data-driven Decision-making Algorithms". In: *arXiv preprint arXiv:2206.14983* (2022).
- [81] Amanda Coston, Anna Kawakami, Haiyi Zhu, Ken Holstein, and Hoda Heidari. "A validity perspective on evaluating the justified use of data-driven decision-making algorithms". In: *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE. 2023, pp. 690–704.
- [82] Amanda Coston, Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. "Counterfactual risk assessments, evaluation, and fairness". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 582–593.
- [83] David A Cotter, Joan M Hermesen, Seth Ovadia, and Reeve Vanneman. "The glass ceiling effect". In: *SF* (2001).
- [84] Council on Social Work Education. *2022 - 2023 Statistics on Social Work Education in the United States*. 2023.
- [85] Bo Cowgill. "The Impact of Algorithms on Judicial Discretion: Evidence from Regression Discontinuities". In: (2018). URL: <http://www.columbia.edu/~bc2656/papers/RecidAlgo.pdf>.
- [86] Kate Crawford. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press, 2021.
- [87] National Council on Crime and Delinquency Children's Research Center. *The Structured Decision Making System for Child Protective Services Policy and Procedures Manual*. <https://www.cdss.ca.gov/Portals/9/SDM%20Policy%20and%20Procedure%20Manual.pdf>. Online; accessed 8-September-2021.
- [88] Tim Dare and Eileen Gambrill. "Ethical Analysis: Predictive Risk Models at Call Screening for Allegheny County". In: (2016). Online; accessed 8-September-2021. URL: https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/Ethical-Analysis-16-ACDHS-26-PredictiveRisk-Package_050119_FINAL-2.pdf.
- [89] Tim Dare and Eileen Gambrill. "Ethical analysis: Predictive risk models at call screening for Allegheny County". In: *Alleghany County Analytics* (2017).
- [90] Scott Davidoff, Min Kyung Lee, Anind K Dey, and John Zimmerman. "Rapidly exploring application design through speed dating". In: *International Conference on Ubiquitous Computing*. Springer. 2007, pp. 429–446.
- [91] Maria De-Arteaga, Artur Dubrawski, and Alexandra Chouldechova. "Learning under selective labels in the presence of expert consistency". In: *arXiv preprint arXiv:1807.00905* (2018).
- [92] Maria De-Arteaga, Artur Dubrawski, and Alexandra Chouldechova. "Leveraging expert consistency to improve algorithmic decision support". In: *arXiv* (2021), pp. 1–33. arXiv: 2101.09648. URL: <http://arxiv.org/abs/2101.09648>.

- [93] Maria De-Arteaga, Artur Dubrawski, and Alexandra Chouldechova. "Leveraging Expert Consistency to Improve Algorithmic Decision Support". In: *arXiv preprint arXiv:2101.09648* (2021).
- [94] Maria De-Arteaga, Artur Dubrawski, and Alexandra Chouldechova. "Leveraging expert consistency to improve algorithmic decision support". In: *arXiv preprint arXiv:2101.09648* (2021).
- [95] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. "A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores". In: *arXiv* (2020), pp. 1–12. ISSN: 23318422.
- [96] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. "A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–12.
- [97] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. "Stakeholder Participation in AI: Beyond" Add Diverse Stakeholders and Stir"". In: *arXiv preprint arXiv:2111.01122* (2021).
- [98] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. "The participatory turn in AI design: Theoretical foundations and the current state of practice". In: *EAAMO'23*. 2023.
- [99] Michael Desmond, Zahra Ashktorab, Werner Geyer, Elizabeth M Daly, Martin Santillan Cooper, Qian Pan, Rahul Nair, Nico Wagner, and Tejaswini Pedapati. "EvalAssist: LLLM-as-a-judge simplified". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. 28. 2025, pp. 29637–29639.
- [100] Alan J Dettlaff and Reiko Boyd. "Racial disproportionality and disparities in the child welfare system: Why do they exist, and what can be done to address them?" In: *The ANNALS of the American Academy of Political and Social Science* 692.1 (2020), pp. 253–274.
- [101] Alan J Dettlaff, Stephanie L Rivaux, Donald J Baumann, John D Fluke, Joan R Rycraft, and Joyce James. "Disentangling substantiation: The influence of race, income, and risk on the substantiation decision in child welfare". In: *Children and Youth Services Review* 33.9 (2011), pp. 1630–1637.
- [102] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. "Algorithm aversion: People erroneously avoid algorithms after seeing them err." In: *Journal of Experimental Psychology: General* 144.1 (2015), p. 114.
- [103] Tawanna R Dillahunt, Jason Lam, Alex Lu, and Earnest Wheeler. "Designing future employment applications for underserved job seekers: a speed dating study". In: *Proceedings of the 2018 Designing Interactive Systems Conference*. 2018, pp. 33–44.
- [104] Kate Donahue, Alexandra Chouldechova, and Krishnaram Kenthapadi. "Human-Algorithm Collaboration: Achieving Complementarity and Avoiding Unfairness". In: *arXiv preprint arXiv:2202.08821* (2022).
- [105] Dorothy Roberts. *Shattered bonds: The color of child welfare*. New York: Basic Books, 2002.
- [106] Paul Dourish, Christopher Lawrence, Tuck Wah Leong, and Greg Wadley. "On being iterated: The affective demands of design participation". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–11.

- [107] Andy Dow, Rob Comber, and John Vines. "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*. 2018, pp. 1–13.
- [108] Brett Drake, Jennifer M Jolley, Paul Lanier, John Fluke, Richard P Barth, and Melissa Jonson-Reid. "Racial bias in child protection? A comparison of competing explanations using national data". In: *Pediatrics* 127.3 (2011), pp. 471–478.
- [109] Julia Dressel and Hany Farid. "The accuracy, fairness, and limits of predicting recidivism". In: *Science advances* 4.1 (2018), eaao5580.
- [110] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. "Fairness through awareness". In: *Proceedings of the 3rd innovations in theoretical computer science conference*. 2012, pp. 214–226.
- [111] Gilan M El Saadawi, Roger Azevedo, Melissa Castine, Velma Payne, Olga Medvedeva, Eugene Tseytlin, Elizabeth Legowski, Drazen Jukic, and Rebecca S Crowley. "Factors affecting feeling-of-knowing in a medical intelligent tutoring system: the role of immediate feedback as a metacognitive scaffold". In: *Advances in Health Sciences Education* 15.1 (2010), pp. 9–30.
- [112] B-N Sanders Elizabeth, Pieter Jan Stappers, et al. "Co-creation and the new landscapes of design". In: *CoDesign* 4.1 (2008), pp. 5–18.
- [113] Glyn Elwyn, Isabelle Scholl, Caroline Tietbohl, Mala Mann, Adrian GK Edwards, Catharine Clay, France Légaré, Trudy van der Weijden, Carmen L Lewis, Richard M Wexler, et al. "Many miles to go...": A systematic review of the implementation of patient decision support interventions into routine clinical practice". In: *BMC MIDM* (2013).
- [114] Paula England. "Emerging theories of care work". In: *ARS* (2005).
- [115] David Freeman Engstrom, Daniel E Ho, Catherine M Sharkey, and Mariano-Florentino Cuéllar. "Government by algorithm: Artificial intelligence in federal administrative agencies". In: *NYU School of Law, Public Law Research Paper* 20-54 (2020).
- [116] Noel Enyedy. "Personalized instruction: New interest, old rhetoric, limited results, and the need for a new direction for computer-mediated learning". In: *NEPC* (2014).
- [117] Raffaele Esposito, Laura Fiorini, Raffaele Limosani, Manuele Bonaccorsi, Alessandro Manzi, Filippo Cavallo, and Paolo Dario. "Supporting active and healthy aging with advanced robotics integrated in smart environment". In: *Optimizing Assistive Technologies for Aging Populations*. 2016.
- [118] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. "Dermatologist-level classification of skin cancer with deep neural networks". In: *nature* 542.7639 (2017), pp. 115–118.
- [119] Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.
- [120] Mark Fabian, Anna Alexandrova, and Yamini Cinamon Nair. "Coproducing Wellbeing Policy: A Theory of Thriving in Financial Hardship". In: *Journal of Happiness Studies* 24.7 (2023), pp. 2309–2330.

- [121] Cathy Rozel Farnworth. "Well-being is a process of becoming: Respondent-led research with organic farmers in Madagascar". In: *Social Indicators Research* 90.1 (2009), pp. 89–106.
- [122] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. "Certifying and removing disparate impact". In: *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 2015, pp. 259–268.
- [123] Shi Feng and Jordan Boyd-Graber. "What can ai do for me? evaluating machine learning interpretations in cooperative play". In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 2019, pp. 229–239.
- [124] Emilia Filippi, Mariasole Banno, and Sandro Trento. "Automation technologies and the risk of substitution of women: Can gender equality in the institutional context reduce the risk?" In: *Technological Forecasting and Social Change* 191 (2023), p. 122528.
- [125] Grace Flores-Robles and Ana P Gantman. "Notions of Care Labor Are Antithetical to Profitable Labor". In: *PWQ* (2023).
- [126] Riccardo Fogliato, Alice Xiang, Daniel Nagin, and Alexandra Chouldechova. *On the Validity of Arrest as a Proxy for Offense : Race and the Likelihood of Arrest for Violent Crimes*. Association for Computing Machinery, 2021. arXiv: [arXiv: 2105.04953v1](https://arxiv.org/abs/2105.04953v1).
- [127] Diana Forsythe. *Studying those who study us: An anthropologist in the world of artificial intelligence*. Stanford University Press, 2001.
- [128] Diana E Forsythe. "The construction of work in artificial intelligence". In: *ST&HV* (1993).
- [129] Michel Foucault. "Discipline and punish". In: *Social Theory Re-Wired*. 2023.
- [130] Sarah E Fox, Samantha Shorey, Esther Y Kang, Dominique Montiel Valle, and Estefania Rodriguez. "Patchwork: the hidden, human labor of AI integration within essential work". In: *CSCW'23* (2023).
- [131] Sarah E Fox, Kiley Sobel, and Daniela K Rosner. "Managerial Visions: Stories of upgrading and maintaining the public restroom with IoT". In: *CHI'19*. 2019.
- [132] Aline Shakti Franzke, Iris Muis, and Mirko Tobias Schäfer. "Data Ethics Decision Aid (DEDA): a dialogical framework for ethical inquiry of AI and data projects in the Netherlands". In: *Ethics and Information Technology* (2021), pp. 1–17.
- [133] Krzysztof Z Gajos and Lena Mamykina. "Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning". In: *27th International Conference on Intelligent User Interfaces*. 2022, pp. 794–806.
- [134] Child Welfare Information Gateway. *Safety and Risk Assessment*. Online; accessed 25-Dec-2021. URL: <https://www.childwelfare.gov/topics/systemwide/assessment/family-assess/safety/>.
- [135] Simret Araya Gebreegziabher, Charles Chiang, Zichu Wang, Zahra Ashktorab, Michelle Brachman, Werner Geyer, Toby Jia-Jun Li, and Diego Gómez-Zarà. "MetricMate: An Interactive Tool for Generating Evaluation Criteria for LLM-as-a-Judge Workflow". In: *Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work*. 2025, pp. 1–18.

- [136] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. "Datasheets for datasets". In: *CACM* 64.12 (2021), pp. 86–92.
- [137] Lara M Genik, C Meghan McMurtry, Paula C Barata, Chantel C Barney, and Stephen P Lewis. "The importance of feasible outcome evaluations: Developing stakeholder-informed outcomes in a randomized controlled trial for children's respite workers receiving pain training". In: *Paediatric and Neonatal Pain* 3.1 (2021), pp. 12–19.
- [138] Marissa Gerchick, Tobi Jegede, Tarak Shah, Ana Gutierrez, Sophie Beiers, Noam Shemtov, Kath Xu, Anjana Samant, and Aaron Horowitz. "The Devil is in the Details: Interrogating Values Embedded in the Allegheny Family Screening Tool". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 2023, pp. 1292–1310.
- [139] Julia Gilman. *Automated Immunity*. 2018. URL: <https://brownpoliticalreview.org/2018/02/automated-immunity/>.
- [140] Jeremy D Goldhaber-Fiebert and Lea Prince. "Impact evaluation of a predictive risk modeling tool for Allegheny county's child welfare office". In: *Pittsburgh: Allegheny County* (2019).
- [141] Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. "The disagreement deconvolution: Bringing machine learning performance metrics in line with reality". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–14.
- [142] Karine Gravel, France Légaré, and Ian D Graham. "Barriers and facilitators to implementing shared decision-making in clinical practice: A systematic review of health professionals' perceptions". In: *Implementation science* (2006).
- [143] Ben Green. "The false promise of risk assessments: epistemic reform and the limits of fairness". In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020, pp. 594–606.
- [144] Ben Green. "The flaws of policies requiring human oversight of government algorithms". In: *Available at SSRN* (2021).
- [145] Ben Green. "The flaws of policies requiring human oversight of government algorithms". In: *Computer Law & Security Review* 45 (2022), p. 105681.
- [146] Ben Green and Yiling Chen. "Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments". In: *Proceedings of Conference on Fairness, Accountability, and Transparency (FAT*)*. ACM, 2019.
- [147] Ben Green and Yiling Chen. "The principles and limits of algorithm-in-the-loop decision making". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (2019). ISSN: 25730142. DOI: [10.1145/3359152](https://doi.org/10.1145/3359152).
- [148] Ben Green and Yiling Chen. "The principles and limits of algorithm-in-the-loop decision making". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (2019), pp. 1–24.
- [149] Luke Guerdan, Amanda Coston, Zhiwei Steven Wu, and Kenneth Holstein. "Ground (less) Truth: A Causal Framework for Proxy Labels in Human-Algorithm Decision-Making". In: *arXiv preprint arXiv:2302.06503* (2023).

- [150] Luke Guerdan, Devansh Saxena, Stevie Chancellor, Zhiwei Steven Wu, and Kenneth Holstein. "Measurement as Bricolage: Examining How Data Scientists Construct Target Variables for Predictive Modeling Tasks". In: *arXiv preprint arXiv:2507.02819* (2025).
- [151] Lillemor RM Hallberg. "The "core category" of grounded theory: Making constant comparisons". In: *International journal of qualitative studies on health and well-being* 1.3 (2006), pp. 141–148.
- [152] Emil J Haller. "Pupil race and elementary school ability grouping: Are teachers biased against Black children?" In: *American Educational Research Journal* 22.4 (1985), pp. 465–483.
- [153] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. "Towards a critical race methodology in algorithmic fairness". In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020, pp. 501–512.
- [154] Donna Haraway. "Situated knowledges: The science question in feminism and the privilege of partial perspective 1". In: *Women, science, and technology*. Routledge, 2013, pp. 455–472.
- [155] Moritz Hardt, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning". In: *Advances in neural information processing systems* 29 (2016), pp. 3315–3323.
- [156] Ron Haskins. "Child welfare financing: What do we fund, how, and what could be improved?" In: *The ANNALS of the American Academy of Political and Social Science* 692.1 (2020), pp. 50–67.
- [157] Wanrong He, Mitchell L Gordon, Lindsay Popowski, and Michael S Bernstein. "Cura: Curation at Social Media Scale". In: *Proceedings of the ACM on Human-Computer Interaction* 7.CSCW2 (2023), pp. 1–33.
- [158] Anne Helsdingen, Tamara Van Gog, and Jeroen Van Merriënboer. "The effects of practice schedule and critical thinking prompts on learning and transfer of a complex judgment task." In: *Journal of Educational Psychology* 103.2 (2011), p. 383.
- [159] Drew Hemment, Mel Woods, Vimla Appadoo, and Lily Bui. "Community Key Performance Indicators (Community KPIs) for the IoT and Smart Cities". In: *Future Everything: Manchester, UK* (2016).
- [160] Patrick Hemmer, Max Schemmer, Niklas Kühl, Michael Vössing, and Gerhard Satzger. "On the Effect of Information Asymmetry in Human-AI Teams". In: *arXiv preprint arXiv:2205.01467* (2022).
- [161] Joy Hendry and Yongmei Wu. *The care of the elderly in Japan*. Routledge, 2004.
- [162] Beverly Hittle, Noma Agbonifo, Rassull Suarez, Kermit G Davis, and Tangela Ballard. "Complexity of occupational exposures for home health-care workers: Nurses vs. home health aides". In: *JNM* (2016).
- [163] Sally Ho and Garance Burke. "An algorithm that screens for child neglect raises concerns". In: *AP News* (2022).
- [164] Sally Ho and Garance Burke. "Child welfare algorithm faces Justice Department scrutiny". In: *AP News* (2023).
- [165] Anna Lauren Hoffmann. "Terms of inclusion: Data, discourse, violence". In: *New Media & Society* 23.12 (2021), pp. 3539–3556.

- [166] Michael J Hogan, Helen Johnston, Benjamin Broome, Claire McMoreland, Jane Walsh, Bryan Smale, Jim Duggan, Jerry Andriessen, Kevin M Leyden, Christine Domegan, et al. "Consulting with citizens in the design of well-being measures and policies: Lessons from a systems science application". In: *Social Indicators Research* 123 (2015), pp. 857–877.
- [167] Wayne Holmes. "The unintended consequences of artificial intelligence and education". In: (2023).
- [168] Kenneth Holstein and Vincent Aleven. "Designing for human–AI complementarity in K-12 education". In: *AI Magazine* 43.2 (2022), pp. 239–248.
- [169] Kenneth Holstein and Vincent Aleven. "Designing for human-AI complementarity in K-12 education". In: *arXiv preprint arXiv:2104.01266* (2021).
- [170] Kenneth Holstein and Vincent Aleven. "Designing for human-AI complementarity in K-12 education". In: *arXiv preprint arXiv:2104.01266* (2021).
- [171] Kenneth Holstein, Vincent Aleven, and Nikol Rummel. "A conceptual framework for human–AI hybrid adaptivity in education". In: *International Conference on Artificial Intelligence in Education*. Springer. 2020, pp. 240–254.
- [172] Kenneth Holstein, Maria De-Arteaga, Lakshmi Tumati, and Yanghui Cheng. "Toward supporting perceptual complementarity in human-AI collaboration via reflection on unobservables". In: *Proceedings of the ACM on Human-Computer Interaction* 7.CSCW1 (2023), pp. 1–20.
- [173] Kenneth Holstein and Shayan Doroudi. "Fairness and equity in learning analytics systems (FairLAK)". In: *Companion proceedings of the ninth international learning analytics & knowledge conference (LAK 2019)*. 2019, pp. 1–2.
- [174] Kenneth Holstein, Erik Harpstead, Rebecca Gulotta, and Jodi Forlizzi. "Replay Enactments: Exploring possible futures through historical data". In: *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 2020, pp. 1607–1618.
- [175] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. "Co-designing a real-time classroom orchestration tool to support teacher-AI complementarity." In: *Grantee Submission* (2019).
- [176] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. "Co-designing a real-time classroom orchestration tool to support teacher–AI complementarity". In: *Journal of Learning Analytics* 6.2 (2019), pp. 27–52.
- [177] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. "Designing for complementarity: Teacher and student needs for orchestration support in AI-enhanced classrooms". In: *International Conference on Artificial Intelligence in Education*. Springer. 2019, pp. 157–171.
- [178] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. "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*. 2017, pp. 257–266.
- [179] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. "Student learning benefits of a mixed-reality teacher awareness tool in AI-enhanced classrooms". In: *International Conference on Artificial Intelligence in Education*. 2018, pp. 154–168. ISBN: 9783319938431.

- [180] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. "Student learning benefits of a mixed-reality teacher awareness tool in AI-enhanced classrooms". In: *International conference on artificial intelligence in education*. Springer. 2018, pp. 154–168.
- [181] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. "Improving fairness in machine learning systems: What do industry practitioners need?" In: *Proceedings of the 2019 CHI conference on human factors in computing systems*. 2019, pp. 1–16.
- [182] Naja Holten Møller, Irina Shklovski, and Thomas T. Hildebrandt. "Shifting concepts of value: Designing algorithmic decision-support systems for public services". In: *NordiCHI* (2020), pp. 1–12. DOI: [10.1145/3419249.3420149](https://doi.org/10.1145/3419249.3420149).
- [183] Naja Holten Møller, Irina Shklovski, and Thomas T Hildebrandt. "Shifting concepts of value: Designing algorithmic decision-support systems for public services". In: *NordiCHI'20*. 2020.
- [184] Karen Holtzblatt and Hugh Beyer. *Contextual design: defining customer-centered systems*. Elsevier, 1997.
- [185] Jiseung Hong, Grace Byun, Seungone Kim, Kai Shu, and Jinho D Choi. "Measuring sycophancy of language models in multi-turn dialogues". In: *arXiv preprint arXiv:2505.23840* (2025).
- [186] White House. *Fact sheet: President Biden issues executive order on safe, secure, and trustworthy artificial intelligence*. 2023.
- [187] Ruthanne Huising, Kasper Trolle Elmholdt, and Elina Mäkinen. "Introduction: The Changing Constitution and Ecology of Expertise". In: *Research in the Sociology of Organizations* (2025).
- [188] Jessica Hullman, Paul Resnick, and Eytan Adar. "Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering". In: *PloS one* 10.11 (2015), e0142444.
- [189] Allegheny County Department of Human Services. *Allegheny Family Screening Tool, Frequently-Asked Questions | Updated August 2018*. https://www.alleghenycountyanalytics.us/wp-content/uploads/2018/10/17-ACDHS-11_AFST_102518.pdf. Online; accessed 8-September-2021.
- [190] Allegheny County Department of Human Services. "Impact Evaluation Summary of the Allegheny Family Screening Too". In: *Pittsburgh: Allegheny County* (2019).
- [191] Allegheny County Department of Human Services. *The Allegheny Family Screening Tool*. <https://www.alleghenycounty.us/Human-Services/News-Events/Accomplishments/Allegheny-Family-Screening-Tool.aspx>. Online; accessed 7-September-2021.
- [192] (DHS) Allegheny County Department of Human Services. *DHS response to Automated Inequality by Virginia Eubanks*. URL: <https://www.alleghenycounty.us/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=6442461672>.
- [193] Dan Hurley. "Can an Algorithm Tell When Kids Are in Danger?" In: *New York Times Magazine* (). Online; accessed 8-September-2021. URL: [nytimes.com/2018/01/02/magazine/can-an-algorithm-tell-when-kids-are-in-danger.html](https://www.nytimes.com/2018/01/02/magazine/can-an-algorithm-tell-when-kids-are-in-danger.html).
- [194] Dan Hurley. "Can an algorithm tell when kids are in danger". In: *New York Times* 2 (2018).

- [195] SAS Institute. *Analytics for Child Well-Being*. https://www.sas.com/en_us/software/analytics-for-child-well-being.html. Online; accessed 12-December-2021.
- [196] Sheriff's Justice Institute. *Central Bond Court Report*. 2016. URL: https://www.chicagoreader.com/pdf/20161026/Sheriff_s-Justice-Institute-Central-BondCourt-Study-070616.pdf.
- [197] Investment Dialogue for Australia's Children. *The Nest Technical Document 3*. Accessed: 2025-01-22. 2024. URL: <https://www.investmentdialogue.org.au/wp-content/uploads/2024/10/The-Nest-technical-document-3.pdf>.
- [198] Lilly C Irani and M Six Silberman. "Turkopticon: Interrupting worker invisibility in amazon mechanical turk". In: *CHI'13*. 2013.
- [199] D. Jackson and G. Marx. *Data mining program designed to predict child abuse proves unreliable, DCFS says*. 2017. URL: <http://www.chicagotribune.com/news/watchdog/ct-dcfs-eckerd-met-20171206-story.html>.
- [200] David Jackson and Gary Marx. "Data mining program designed to predict child abuse proves unreliable, DCFS says". In: *Chicago Tribune* (). Online; accessed 6-January-2022. URL: <https://www.chicagotribune.com/investigations/ct-dcfs-eckerd-met-20171206-story.html>.
- [201] Abigail Z Jacobs and Hanna Wallach. "Measurement and fairness". In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021, pp. 375–385.
- [202] Abigail Z Jacobs and Hanna Wallach. "Measurement and fairness". In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021, pp. 375–385.
- [203] Nari Johnson, Deepthi Sudharsan, Samantha Dalal, Theo Holroyd, Anja Thieme, Hoda Heidari, Daniela Massiceti, Jennifer Wortman Vaughan, Cecily Morrison, et al. "Evaluating AI-Generated Images of Cultural Artifacts with Community-Informed Rubrics". In: *arXiv preprint arXiv:2604.02406* (2026).
- [204] Will Johnson. *Effectiveness of California's child welfare Structured Decision-Making (SDM) model: A Prospective study of the validity of the California family risk assessment*. 2004.
- [205] Andreas Møller Jørgensen and Maria Appel Nissen. "Making sense of decision support systems: Rationales, translations and potentials for critical reflections on the reality of child protection". In: *BD&S* (2022).
- [206] NPR June. "Oregon is dropping an artificial intelligence tool used in child welfare system". In: *Link: https://www.npr.org/2022/06/02/1102661376/oregon-drops-artificial-intelligence-child-abuse-cases* (2022).
- [207] Dee Jupp, Sohel Ibn Ali, and Carlos Barahona. "Measuring empowerment? Ask them". In: *Sida Studies in Evaluation* 1 (2010), pp. 7–99.
- [208] Ekaterina Jussupow, Kai Spohrer, Armin Heinzl, and Joshua Gawlitza. "Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence". In: *Information Systems Research* (2021).
- [209] Daniel Kahneman, Olivier Sibony, and Cass R Sunstein. *Noise: A flaw in human judgment*. Little, Brown, 2021.

- [210] Sarah J Kaka, Joshua Littenberg-Tobias, Taylor Kessner, Anthony Tuf Francis, Katrina Kennett, G Marvez, and Justin Reich. "Digital simulations as approximations of practice: Preparing preservice teachers to facilitate whole-class discussions of controversial issues". In: *Journal of Technology and Teacher Education* 29.1 (2021), pp. 67–90.
- [211] Alex Kale, Matthew Kay, and Jessica Hullman. "Visual reasoning strategies for effect size judgments and decisions". In: *IEEE Transactions on Visualization and Computer Graphics* 27.2 (2020), pp. 272–282.
- [212] Nathan Kallus and Angela Zhou. "Residual unfairness in fair machine learning from prejudiced data". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 2439–2448.
- [213] Ece Kamar. "Directions in hybrid intelligence: Complementing AI systems with human intelligence". In: *International Joint Conference on Artificial Intelligence (IJCAI)* (2016), pp. 4070–4073.
- [214] Esther Y Kang and Sarah E Fox. "Stories from the Frontline: Recuperating Essential Worker Accounts of AI Integration". In: *Designing Interactive Systems Conference*. 2022, pp. 58–70.
- [215] Fadhilah M Karimah and Inaki M Hakim. "Designing Strategies for Inbound Logistics Performance Indicator Improvement in the Indonesian Mining Industry". In: *Proceedings of the 2020 International Conference on Engineering and Information Technology for Sustainable Industry*. 2020, pp. 1–7.
- [216] Anna Kawakami, Shreya Chowdhary, Shamsi T Iqbal, Q Vera Liao, Alexandra Olteanu, Jina Suh, and Koustuv Saha. "Sensing Wellbeing in the Workplace, Why and For Whom? Envisioning Impacts with Organizational Stakeholders". In: *CSCW'23* (2023).
- [217] Anna Kawakami, Amanda Coston, Hoda Heidari, Kenneth Holstein, and Haiyi Zhu. "Studying Up Public Sector AI: How Networks of Power Relations Shape Agency Decisions Around AI Design and Use". In: (2023).
- [218] Anna Kawakami, Amanda Coston, Hoda Heidari, Kenneth Holstein, and Haiyi Zhu. "Studying Up Public Sector AI: How Networks of Power Relations Shape Agency Decisions Around AI Design and Use". In: *CSCW'24* (2024).
- [219] Anna Kawakami, Amanda Coston, Haiyi Zhu, Hoda Heidari, and Kenneth Holstein. "The Situate AI Guidebook: Co-Designing a Toolkit to Support Multi-Stakeholder, Early-stage Deliberations Around Public Sector AI Proposals". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 2024, pp. 1–22.
- [220] Anna Kawakami, Luke Guerdan, Yanghuidi Cheng, Kate Glazko, Matthew Lee, Scott Carter, Nikos Arechiga, Haiyi Zhu, and Kenneth Holstein. "Training towards critical use: Learning to situate AI predictions relative to human knowledge". In: *Proceedings of The ACM Collective Intelligence Conference*. 2023, pp. 63–78.
- [221] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. "Improving human-AI partnerships in child welfare: understanding worker practices, challenges, and desires for algorithmic decision support". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–18.

- [222] Anna Kawakami, Venkatesh Sivaraman, Logan Stapleton, Hao-Fei Cheng, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. ““Why do I care what’s similar?” Probing challenges in AI-assisted child welfare decision-making through worker-AI interface design concepts”. In: *Proceedings of the 2022 ACM Designing Interactive Systems Conference*. 2022, pp. 454–470.
- [223] Anna Kawakami, Jordan Taylor, Sarah Fox, Haiyi Zhu, and Kenneth Holstein. “AI failure loops in devalued work: The confluence of overconfidence in AI and underconfidence in worker expertise”. In: *Big Data & Society* 13.1 (2026), p. 20539517261424164.
- [224] Anna Kawakami, Haiyi Zhu, and Kenneth Holstein. “AI Measurement as a Collaborative Design Practice: A Multidisciplinary Synthesis and Future Directions”. In: *Available at SSRN* 6276543 (2026).
- [225] Matthew Kay, Tara Kola, Jessica Hullman, and Sean A. Munson. “When (ish) is My Bus? User-centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems”. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2016, pp. 5092–5103. ISBN: 9781450333627. DOI: [10.1145/2858036.2858558](https://doi.org/10.1145/2858036.2858558).
- [226] Emily Keddell. “Algorithmic Justice in Child Protection: Statistical Fairness, Social Justice and the Implications for Practice”. In: *Social Sciences* (2019).
- [227] Patrick Gage Kelley, Steven Rousso-Schindler, Renee Shelby, Kurt Thomas, and Allison Woodruff. “How Generative AI Empowers Attackers and Defenders Across the Trust & Safety Landscape”. In: *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 2026, pp. 1–21.
- [228] Connor M Kerns, Diana L Robins, Paul T Shattuck, Craig J Newschaffer, and Steven J Berkowitz. “Expert consensus regarding indicators of a traumatic reaction in autistic youth: A Delphi survey”. In: *Journal of Child Psychology and Psychiatry* 64.1 (2023), pp. 50–58.
- [229] Gavin Kerrigan, Padhraic Smyth, and Mark Steyvers. “Combining Human Predictions with Model Probabilities via Confusion Matrices and Calibration”. In: *Advances in Neural Information Processing Systems* 34 (2021).
- [230] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. “Avoiding discrimination through causal reasoning”. In: *arXiv preprint arXiv:1706.02744* (2017).
- [231] Dae Hyun Kim, Hyungyu Shin, Shakhnozakhon Yadgarova, Jinho Son, Hariharan Subramonyam, and Juho Kim. “AINeedsPlanner: A Workbook to Support Effective Collaboration Between AI Experts and Clients”. In: *Proceedings of the 2024 ACM Designing Interactive Systems Conference*. 2024, pp. 728–742.
- [232] Hyunil Kim, Christopher Wildeman, Melissa Jonson-Reid, and Brett Drake. “Lifetime Prevalence of Investigating Child Maltreatment Among US Children”. In: *American Journal of Public Health* 107.2 (2017), pp. 274–280. DOI: [10.2105/AJPH.2016.303545](https://doi.org/10.2105/AJPH.2016.303545).
- [233] Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M Bean, Katerina Margatina, Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, et al. “The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 105236–105344.

- [234] Rob Kitchin and Tracey Lauriault. "Towards critical data studies: Charting and unpacking data assemblages and their work". In: (2014).
- [235] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. "Human decisions and machine predictions". In: *The quarterly journal of economics* 133.1 (2018), pp. 237–293.
- [236] Janin Koch and Antti Oulasvirta. "Group cognition and collaborative AI". In: *Human and Machine Learning*. Springer, 2018, pp. 293–312.
- [237] Amanda Kocon. *Is Teaching Undervalued Because It's "Women's Work"?* 2017. URL: <https://tntp.org/blog/is-teaching-undervalued-because-its-womens-work/>.
- [238] Kenneth R Koedinger and Vincent Aleven. "Exploring the assistance dilemma in experiments with cognitive tutors". In: *Educational Psychology Review* 19.3 (2007), pp. 239–264.
- [239] Kenneth R Koedinger, Albert T Corbett, and Charles Perfetti. "The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning". In: *Cognitive science* 36.5 (2012), pp. 757–798.
- [240] PM Krafft, Meg Young, Michael Katell, Jennifer E Lee, Shankar Narayan, Micah Epstein, Dharma Dailey, Bernease Herman, Aaron Tam, Vivian Guetler, et al. "An action-oriented AI policy toolkit for technology audits by community advocates and activists". In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021, pp. 772–781.
- [241] Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. "Supporting reflective public thought with considerit". In: *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. 2012, pp. 265–274.
- [242] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. "Tell me more? The effects of mental model soundness on personalizing an intelligent agent". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2012, pp. 1–10.
- [243] Bogdan Kulynych, David Madras, Smitha Milli, Inioluwa Deborah Raji, Angela Zhou, and Richard Zemel. "Participatory approaches to machine learning". In: *International Conference on Machine Learning Workshop*. Vol. 7. 2020.
- [244] Elizabeth Fong-Chy Kuo, Jacklyn Cho, Iredia Olaye, Diana Delgado, Nicola Dell, and Madeline R Sterling. "Understanding the Technological Landscape of Home Health Aides: Scoping Literature Review and a Landscape Analysis of Existing mHealth Apps". In: *JMIR* (2022).
- [245] Tzu-Sheng Kuo, Aaron Lee Halfaker, Zirui Cheng, Jiwoo Kim, Meng-Hsin Wu, Tongshuang Wu, Kenneth Holstein, and Haiyi Zhu. "Wikibench: Community-driven data curation for ai evaluation on wikipedia". In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 2024, pp. 1–24.
- [246] Tzu-Sheng Kuo, Sophia Liu, Quan Ze Chen, Joseph Seering, Amy X Zhang, Haiyi Zhu, and Kenneth Holstein. "Botender: Supporting Communities in Collaboratively Designing AI Agents through Case-Based Provocations". In: *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 2026, pp. 1–31.

- [247] Tzu-Sheng Kuo, Hong Shen, Jisoo Geum, Nev Jones, Jason I Hong, Haiyi Zhu, and Kenneth Holstein. "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*. 2023, pp. 1–17.
- [248] Maciej Kuziemski and Gianluca Misuraca. "AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings". In: *Telecommunications policy* 44.6 (2020), p. 101976.
- [249] Kwadwo Kyeremanteng and Gianni D'Egidio. "Why process quality measures may be more valuable than outcome measures in critical care patients". In: *Biology and Medicine* 7.2 (2015), p. 1.
- [250] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. "Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies". In: *arXiv preprint arXiv:2112.11471* (2021).
- [251] Vivian Lai, Han Liu, and Chenhao Tan. "" Why is' Chicago' deceptiv?" Towards Building Model-Driven Tutorials for Humans". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–13.
- [252] Vivian Lai and Chenhao Tan. "On human predictions with explanations and predictions of machine learning models: A case study on deception detection". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 2019, pp. 29–38.
- [253] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. "Building machines that learn and think like people". In: *Behavioral and brain sciences* 40 (2017).
- [254] Michelle S Lam, Fred Hohman, Dominik Moritz, Jeffrey P Bigham, Kenneth Holstein, and Mary Beth Kery. "AI Policy Projector: Grounding LLM Policy Design in Iterative Mapmaking". In: *arXiv preprint arXiv:2409.18203* (2024).
- [255] Michelle S Lam, Zixian Ma, Anne Li, Izequiel Freitas, Dakuo Wang, James A Landay, and Michael S Bernstein. "Model sketching: centering concepts in early-stage machine learning model design". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–24.
- [256] Sarah Lebovitz, Natalia Levina, and Hila Lifshitz-Assaf. "IS AI GROUND TRUTH REALLY TRUE? THE DANGERS OF TRAINING AND EVALUATING AI TOOLS BASED ON EXPERTS' KNOW-WHAT." In: *MIS quarterly* (2021).
- [257] Sarah Lebovitz, Hila Lifshitz-Assaf, and Natalia Levina. "To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis". In: *Organization science* (2022).
- [258] John D Lee and Katrina A See. "Trust in automation: Designing for appropriate reliance". In: *Human factors* 46.1 (2004), pp. 50–80.
- [259] Min Kyung Lee, Ishan Nigam, Angie Zhang, Joel Afriyie, Zhizhen Qin, and Sicun Gao. "Participatory algorithmic management: Elicitation methods for worker well-being models". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 2021, pp. 715–726.
- [260] Victoria K Lee and Lasana T Harris. "How social cognition can inform social decision making". In: *Frontiers in neuroscience* 7 (2013), p. 259.

- [261] Karen Levy, Kyla E Chasalow, and Sarah Riley. "Algorithms and Decision-Making in the Public Sector". In: *Annual Review of Law and Social Science* 17 (2021), pp. 1–38.
- [262] Karen Levy, Kyla E Chasalow, and Sarah Riley. "Algorithms and decision-making in the public sector". In: *Annual Review of Law and Social Science* 17 (2021), pp. 309–334.
- [263] Belinda Z Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. "Eliciting human preferences with language models". In: *arXiv preprint arXiv:2310.11589* (2023).
- [264] Q Vera Liao, Daniel Gruen, and Sarah Miller. "Questioning the AI: informing design practices for explainable AI user experiences". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–15.
- [265] Sanjiv Lingayah and Florian Sommer. "Communities count: the LITMUS Test". In: *Reflecting Community Indicators in the London Borough of Southwark*. New Economics Foundation, London (2001).
- [266] Michael Lipsky. *Street Level Bureaucracy*. Russell Sage Foundation, 1980. ISBN: 9780871545244. URL: <http://www.jstor.org/stable/10.7758/9781610447713>.
- [267] Zachary C Lipton. "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." In: *Queue* 16.3 (2018), pp. 31–57.
- [268] Yu Lu Liu, Su Lin Blodgett, Jackie Chi Kit Cheung, Q Vera Liao, Alexandra Olteanu, and Ziang Xiao. "ECBD: Evidence-Centered Benchmark Design for NLP". In: *arXiv preprint arXiv:2406.08723* (2024).
- [269] Michele Loi and Matthias Spielkamp. "Towards accountability in the use of artificial intelligence for public administrations". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 2021, pp. 757–766.
- [270] Simone Lombardini, Kimberly Bowman, and Rosa Garwood. *A'how to'guide to measuring women's empowerment: Sharing experience from Oxfam's impact evaluations*. Oxfam GB, 2017.
- [271] Zhuoran Lu and Ming Yin. "Human reliance on machine learning models when performance feedback is limited: Heuristics and risks". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–16.
- [272] Kristian Lum, David B. Dunson, and James Johndrow. "Closer than they appear: A Bayesian perspective on individual-level heterogeneity in risk assessment". In: *arXiv* (2021). arXiv: 2102.01135. URL: <http://arxiv.org/abs/2102.01135>.
- [273] Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery". In: *Nature biomedical engineering* 2.10 (2018), pp. 749–760.
- [274] Lin Luo, Yuri Nakao, Mathieu Chollet, Hiroya Inakoshi, and Simone Stumpf. "EARN Fairness: Explaining, Asking, Reviewing and Negotiating Artificial Intelligence Fairness Metrics Among Stakeholders". In: *arXiv preprint arXiv:2407.11442* (2024).

- [275] Emma Lurie and Deirdre K Mulligan. "Crowdworkers are not judges: Rethinking crowdsourced vignette studies as a risk assessment evaluation technique". In: *Proceedings of the Workshop on Fair and Responsible AI at CHI 2020* (2020).
- [276] Claude Lvi-Strauss. *The savage mind*. University of Chicago Press, 1966.
- [277] Roger Mac Ginty. "Indicators+: A proposal for everyday peace indicators". In: *Evaluation and Program Planning* 36.1 (2013), pp. 56–63.
- [278] Roger Mac Ginty and Pamina Firchow. "Everyday Peace Indicators: Capturing local voices through surveys". In: *Shared Space: A Research Journal on Peace, Conflict and Community Relations in Northern Ireland* 18 (2014), pp. 33–39.
- [279] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. "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*. 2020, pp. 1–14.
- [280] David Madras, Toniann Pitassi, and Richard Zemel. "Predict responsibly: improving fairness and accuracy by learning to defer". In: *arXiv preprint arXiv:1711.06664* (2017).
- [281] Donald Martin Jr, Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S Isaac. "Participatory problem formulation for fairer machine learning through community based system dynamics". In: *arXiv preprint arXiv:2005.07572* (2020).
- [282] Alexandra Mateescu and Madeleine Elish. "AI in context: The labor of integrating new technologies". In: (2019).
- [283] J Nathan Matias and Megan Price. "How public involvement can improve the science of AI". In: *Proceedings of the National Academy of Sciences* 122.48 (2025), e242111122.
- [284] Shefeh Prisilia Mbuy and Marco Ortolani. "Algorithmic Impact Assessment for an Ethical Use of AI in SMEs". In: 2022. DOI: [10.14236/ewic/HCI2022.34](https://scienceopen.com/hosted-document?doi=10.14236/ewic/HCI2022.34). URL: <https://scienceopen.com/hosted-document?doi=10.14236/ewic/HCI2022.34>.
- [285] William E McAuliffe. "Measuring the quality of medical care: process versus outcome". In: *The Milbank Memorial Fund Quarterly. Health and Society* (1979), pp. 118–152.
- [286] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. "Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (2019), pp. 1–23.
- [287] Aditya Krishna Menon and Robert C Williamson. "The cost of fairness in binary classification". In: *Conference on Fairness, Accountability and Transparency*. PMLR. 2018, pp. 107–118.
- [288] Milagros Miceli, Julian Posada, and Tianling Yang. "Studying up machine learning data: Why talk about bias when we mean power?" In: *Proceedings of the ACM on Human-Computer Interaction* 6.GROUP (2022), pp. 1–14.
- [289] Ariela Migda. *The Legacies of Slavery and Jim Crow Live on With Exclusion of Home Health Care Workers from Fair Labor Laws*. 2015.

- [290] Joy Ming, Elizabeth Kuo, Katie Go, Emily Tseng, John Kallas, Aditya Vashistha, Madeline Sterling, and Nicola Dell. "' I Go Beyond and Beyond" Examining the Invisible Work of Home Health Aides". In: *Proceedings of the ACM on Human-Computer Interaction* 7.CSCW1 (2023), pp. 1–21.
- [291] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. "Model cards for model reporting". In: *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 220–229.
- [292] Sophie Mitra, Kris Jones, Brandon Vick, David Brown, Eileen McGinn, and Mary Jane Alexander. "Implementing a multidimensional poverty measure using mixed methods and a participatory framework". In: *Social Indicators Research* 110 (2013), pp. 1061–1081.
- [293] Clara Mosquera-Lopez, Sos Agaian, Alejandro Velez-Hoyos, and Ian Thompson. "Computer-aided prostate cancer diagnosis from digitized histopathology: a review on texture-based systems". In: *IEEE reviews in biomedical engineering* 8 (2014), pp. 98–113.
- [294] Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. "Assembling accountability: algorithmic impact assessment for the public interest". In: *Available at SSRN* 3877437 (2021).
- [295] Hussein Mozannar, Arvind Satyanarayan, and David Sontag. "Teaching Humans When To Defer to a Classifier via Exemplars". In: *arXiv preprint arXiv:2111.11297* (2021).
- [296] Sendhil Mullainathan and Ziad Obermeyer. *A machine learning approach to low-value health care: wasted tests, missed heart attacks and mis-predictions*. National Bureau of Economic Research, 2019.
- [297] Sendhil Mullainathan and Ziad Obermeyer. "On the inequity of predicting a while hoping for B". In: *AEA Papers and Proceedings*. Vol. 111. 2021, pp. 37–42.
- [298] Michael J Muller and Sarah Kuhn. "Participatory design". In: *CACM* (1993).
- [299] J Murray. "Public sector vs. private sector: What's the difference". In: *The Balance Small Business* (2021).
- [300] Rashid Mushkani, Shravan Nayak, Hugo Berard, Allison Cohen, Shin Koseki, and Hadrien Bertrand. "LIVS: A pluralistic alignment dataset for inclusive public spaces". In: *arXiv preprint arXiv:2503.01894* (2025).
- [301] Toni Naccarato. "Child welfare informatics: A proposed subspecialty for social work". In: *Children and Youth Services Review* 32.12 (2010), pp. 1729–1734.
- [302] Laura Nader. "Up the anthropologist: Perspectives gained from studying up." In: (1972).
- [303] Arvind Narayanan. "How to recognize AI snake oil". In: *Arthur Miller Lecture on Science and Ethics* (2019).
- [304] Judge Michael Nash. *EXAMINATION OF USING STRUCTURED DECISION MAKING AND PREDICTIVE ANALYTICS IN ASSESSING SAFETY AND RISK IN CHILD WELFARE (ITEM NO. 49-A, AGENDA OF SEPTEMBER 20, 2016)*. Online; accessed 6-January-2022. 2017. URL: [http://ocp.lacounty.gov/Portals/0CP/PDF/Reports/%20and%20Communication/Safety%20and%20Risk%20Assessment/SDM%20and%20Predictive%20Analytics%20Report%20\(Risk%20Assessment%20Tools\)%20\(May%202017\).pdf?ver=2018-10-24-083910-100](http://ocp.lacounty.gov/Portals/0CP/PDF/Reports/%20and%20Communication/Safety%20and%20Risk%20Assessment/SDM%20and%20Predictive%20Analytics%20Report%20(Risk%20Assessment%20Tools)%20(May%202017).pdf?ver=2018-10-24-083910-100).

- [305] Jerome Ng and Jeff Harrison. “Key performance indicators for clinical pharmacy services in New Zealand public hospitals: stakeholder perspectives”. In: *Journal of Pharmaceutical Health Services Research* 1.2 (2010), pp. 75–84.
- [306] Laura Nissen. “Social work and the future in a post-Covid 19 world: A foresight lens and a call to action for the profession”. In: *Journal of Technology in Human Services* 38.4 (2020), pp. 309–330.
- [307] NIST. *NIST AI RMF Playbook*. 2023. URL: https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook.
- [308] NIST. *NIST Risk Management Framework*. 2023. URL: <https://csrc.nist.gov/projects/risk-management/about-rmf>.
- [309] Ben P North, Tim Ellis, James Bron, Toby G Knowles, and James F Turnbull. “The Use of Stakeholder Focus Groups to Identify Indicators for the On-Farm Assessment of Trout Welfare”. In: *Fish welfare* (2008), pp. 243–267.
- [310] Ziad Obermeyer and Ezekiel J Emanuel. “Predicting the future—big data, machine learning, and clinical medicine”. In: *The New England journal of medicine* 375.13 (2016), p. 1216.
- [311] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. “Dissecting racial bias in an algorithm used to manage the health of populations”. In: *Science* 366.6464 (2019), pp. 447–453.
- [312] Wanda J Orlikowski. “Learning from notes: Organizational issues in groupware implementation”. In: *CSCW’92*. 1992.
- [313] Paul Osterman. “Improving job quality for direct care workers”. In: *EDQ* (2019).
- [314] Paul Osterman. “Improving Long-Term Care by Finally Respecting Home-Care Aides”. In: *HCR* 48 (2018), S67–S70.
- [315] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, Sebastian Nowozin, Joshua V Dillon, Balaji Lakshminarayanan, and Jasper Snoek. “Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift”. In: *Advances in Neural Information Processing Systems* NeurIPS (2019).
- [316] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. “Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering”. In: *Conference on health, inference, and learning*. PMLR. 2022, pp. 248–260.
- [317] Laura E Panattoni, Rhema Vaithianathan, Toni Ashton, and Geraint H Lewis. “Predictive risk modelling in health: options for New Zealand and Australia”. In: *Australian Health Review* 35.1 (2011), pp. 45–51.
- [318] Tina M Park, Adriana Alvarado Garcia, Juana Catalina Becerra Sandoval, Jiyoo Chang, Bill Curtis-Davidson, Remi Denton, Seeta Peña Gangadharan, Lara Groves, Jeremy Holland, Kenneth Holstein, et al. “Practicing Inclusivity in AI: Stakeholder Engagement Policy in Action”. In: *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*. 2024, pp. 748–751.
- [319] Samir Passi and Solon Barocas. “Problem formulation and fairness”. In: *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 39–48.
- [320] Samir Passi and Mihaela Vorvoreanu. “Overreliance on AI: Literature review”. In: (2022).

- [321] Bhavik N Patel, Louis Rosenberg, Gregg Willcox, David Baltaxe, Mimi Lyons, Jeremy Irvin, Pranav Rajpurkar, Timothy Amrhein, Rajan Gupta, Safwan Halabi, Curtis Langlotz, Edward Lo, Joseph Mammarappallil, A J Mariano, Geoffrey Riley, Jayne Seekins, Luyao Shen, Evan Zucker, and Matthew P Lungren. "Human-machine partnership with artificial intelligence for chest radiograph diagnosis". In: *npj Digital Medicine* (2019). ISSN: 2398-6352. DOI: [10.1038/s41746-019-0189-7](https://doi.org/10.1038/s41746-019-0189-7). URL: <http://dx.doi.org/10.1038/s41746-019-0189-7>.
- [322] Bhavik N Patel, Louis Rosenberg, Gregg Willcox, David Baltaxe, Mimi Lyons, Jeremy Irvin, Pranav Rajpurkar, Timothy Amrhein, Rajan Gupta, Safwan Halabi, et al. "Human-machine partnership with artificial intelligence for chest radiograph diagnosis". In: *NPJ digital medicine* 2.1 (2019), pp. 1–10.
- [323] Andi Peng, Besmira Nushi, Emre Kıcıman, Kori Inkpen, Siddharth Suri, and Ece Kamar. "What you see is what you get? the impact of representation criteria on human bias in hiring". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 7. 1. 2019, pp. 125–134.
- [324] Walt L Perry. *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation, 2013.
- [325] Marcus Persson, David Redmalm, and Clara Iversen. "Caregivers' use of robots and their effect on work environment—a scoping review". In: *JTHS* (2022).
- [326] Jennifer Pierre, Roderic Crooks, Morgan Currie, Britt Paris, and Irene Pasquetto. "Getting Ourselves Together: Data-centered participatory design research & epistemic burden". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–11.
- [327] Sofie Pilemalm. "Participatory design in emerging civic engagement initiatives in the new public sector: Applying PD concepts in resource-scarce organizations". In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 25.1 (2018), pp. 1–26.
- [328] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. "Manipulating and measuring model interpretability". In: (2021). ISSN: 23318422. arXiv: [1802.07810](https://arxiv.org/abs/1802.07810).
- [329] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. "Manipulating and measuring model interpretability". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–52.
- [330] Allied Media Projects. *A People's Guide To Tech*. 2018. URL: <https://alliedmedia.org/resources/peoples-guide-to-ai>.
- [331] Dasha Pruss. "Ghosting the Machine: Judicial Resistance to a Recidivism Risk Assessment Instrument". In: *FAccT'23*. 2023.
- [332] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. "Data cards: Purposeful and transparent dataset documentation for responsible ai". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022, pp. 1776–1826.
- [333] Natalie Naïri Quinn and Simone Lombardini. "The Participatory Index of Women's Empowerment: development and an application in Tunisia". In: *Oxford Development Studies* 52.1 (2024), pp. 54–73.

- [334] Inioluwa Deborah Raji and Joy Buolamwini. "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019, pp. 429–435.
- [335] Inioluwa Deborah Raji, Roxana Daneshjou, and Emily Alsentzer. *It's time to bench the medical exam benchmark*. 2025.
- [336] Inioluwa Deborah Raji, I Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. "The fallacy of AI functionality". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022, pp. 959–972.
- [337] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. "Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices". In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW1 (2021), pp. 1–23.
- [338] Charvi Rastogi, Liu Leqi, Kenneth Holstein, and Hoda Heidari. "A unifying framework for combining complementary strengths of humans and ML toward better predictive decision-making". In: *arXiv preprint arXiv:2204.10806* (2022).
- [339] Helene Friis Ratner and Kasper Elmholdt. "Algorithmic constructions of risk: Anticipating uncertain futures in child protection services". In: *BD&S* (2023).
- [340] Jude Rayan, Dhruv Kanetkar, Yifan Gong, Yuewen Yang, Srishti Palani, Haijun Xia, and Steven P Dow. "Exploring the potential for generative ai-based conversational cues for real-time collaborative ideation". In: *Proceedings of the 16th Conference on Creativity & Cognition*. 2024, pp. 117–131.
- [341] Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker. "Algorithmic Impact Assessments: A Practical Framework for Public Agency". In: *AI Now* (2018).
- [342] John Richards, David Piorkowski, Michael Hind, Stephanie Houde, and Aleksandra Mojsilović. "A methodology for creating AI FactSheets". In: *arXiv preprint arXiv:2006.13796* (2020).
- [343] Naomi Schaefer Riley. *Can Big Data Help Save Abused Kids?* Online; accessed 8-September-2021. 2018. URL: <https://reason.com/2018/01/22/can-big-data-help-save-abused/>.
- [344] Dorothy Roberts. *Torn apart: how the child welfare system destroys black families—and how abolition can build a safer world*. Basic Books, 2022.
- [345] Samantha Robertson, Tonya Nguyen, Cathy Hu, Catherine Albiston, Afshin Nikzad, and Niloufar Salehi. "Expressiveness, Cost, and Collectivism: How the Design of Preference Languages Shapes Participation in Algorithmic Decision-Making". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–16.
- [346] Samantha Robertson, Tonya Nguyen, and Niloufar Salehi. "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*. 2021, pp. 1–14.
- [347] Samantha Robertson and Niloufar Salehi. "What if I don't like any of the choices? The limits of preference elicitation for participatory algorithm design". In: *arXiv preprint arXiv:2007.06718* (2020).

- [348] Keara Rodela, Noelle Wiggins, Kenneth Maes, Teresa Campos-Dominguez, Victoria Adewumi, Pennie Jewell, and Susan Mayfield-Johnson. "The community health worker (CHW) common indicators project: engaging CHWs in measurement to sustain the profession". In: *Frontiers in public health* 9 (2021), p. 674858.
- [349] Ido Roll and Ruth Wylie. "Evolution and revolution in artificial intelligence in education". In: *IJAIED'16* (2016).
- [350] Henriette Rönsch, Frederike Schiffers, Robert Ofenloch, Elke Weisshaar, Anna-Sophie Buse, Andreas Hansen, Swen Malte John, Ana M Giménez Arnau, David Pesqué, Tove Agner, et al. "Which outcomes should be measured in hand eczema trials? Results from patient interviews and an expert survey". In: *Journal of the European Academy of Dermatology and Venereology* 37.6 (2023), pp. 1199–1206.
- [351] Aymeric Roucher. "Using LLM-as-a-judge for an automated and versatile evaluation". In: *Hugging Face* (). URL: https://huggingface.co/learn/cookbook/llm_judge.
- [352] John L Rury. *Education and Women's Work: Female Schooling and the Division of Labor in Urban America, 1870-1930*. SUNY Series on Women and Work. ERIC, 1991.
- [353] Matthew J. Salganik et al. "Measuring the predictability of life outcomes with a scientific mass collaboration". In: *Proceedings of the National Academy of Sciences* 117.15 (2020), pp. 8398–8403. DOI: [10.1073/pnas.1915006117](https://doi.org/10.1073/pnas.1915006117). eprint: <https://www.pnas.org/doi/pdf/10.1073/pnas.1915006117>. URL: <https://www.pnas.org/doi/abs/10.1073/pnas.1915006117>.
- [354] Anjana Samant, Aaron Horowitz, Kath Xu, and Sophie Beiers. "Family surveillance by algorithm: The rapidly spreading tools few have heard of". In: *American Civil Liberties Union (ACLU)* (2021). URL: https://www.aclu.org/sites/default/files/field_document/2021.09.28a_family_surveillance_by_algorithm.pdf.
- [355] Anjana Samant, Aaron Horowitz, Kath Xu, and Sophie Beiers. *Family surveillance by algorithm: The rapidly spreading tools few have heard of*. American Civil Liberties Union (ACLU)(2021). 2021.
- [356] Elizabeth B-N Sanders and Pieter Jan Stappers. "Co-creation and the new landscapes of design". In: *Co-design* 4.1 (2008), pp. 5–18.
- [357] Mariano Sanz, Ana Carrillo de Albornoz, Conchita Martin, Ian Needleman, and Maurizio S Tonetti. "Multi-stakeholder contribution to the identification of a core outcome set and measurements in implant dentistry (ID-COSM initiative) using the Delphi methodology". In: *Journal of Clinical Periodontology* 50 (2023), pp. 107–121.
- [358] Joaquin Sapien. "Foiled by FOIL: How One City Agency Has Dragged Out a Request for Public Records for Nearly a Year". In: *ProPublica* (2016). URL: <https://www.propublica.org/article/how-city-agency-dragged-out-request-for-public-records-for-nearly-a-year>.
- [359] Devansh Saxena, Karla Badillo-Urquiola, Pamela Wisniewski, and Shion Guha. "A framework of high-stakes algorithmic decision-making for the public sector developed through a case study of child welfare". In: *arXiv* 5.October (2021). arXiv: [arXiv:2107.03487v2](https://arxiv.org/abs/2107.03487v2).

- [360] Devansh Saxena, Karla Badillo-Urquiola, Pamela J Wisniewski, and Shion Guha. "A framework of high-stakes algorithmic decision-making for the public sector developed through a case study of child-welfare". In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2 (2021), pp. 1–41.
- [361] Devansh Saxena, Karla Badillo-Urquiola, Pamela J Wisniewski, and Shion Guha. "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*. 2020, pp. 1–15.
- [362] Devansh Saxena, Erina Seh-Young Moon, Aryan Chaurasia, Yixin Guan, and Shion Guha. "Rethinking" Risk" in Algorithmic Systems Through A Computational Narrative Analysis of Casenotes in Child-Welfare". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–19.
- [363] Devansh Saxena, Seh Young Moon, Dahlia Shehata, and Shion Guha. "Unpacking Invisible Work Practices, Constraints, and Latent Power Relationships in Child Welfare through Casenote Analysis". In: *CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–22.
- [364] China Scherz. "Protecting children, preserving families: Moral conflict and actuarial science in a problem of contemporary governance". In: *PoLAR: Political and Legal Anthropology Review* 34.1 (2011), pp. 33–50.
- [365] Donald A Schön. "Designing as reflective conversation with the materials of a design situation". In: *Knowledge-based systems* 5.1 (1992), pp. 3–14.
- [366] *School of Social Work study confirms human services workers are underpaid*. 2022. URL: <https://socialwork.uw.edu/news/school-social-work-study-confirms-human-services-workers-are-underpaid/>.
- [367] Thomas A Schwandt and Emily F Gates. "Case study methodology". In: *The SAGE handbook of qualitative research* 5 (2018), pp. 600–630.
- [368] James C Scott. *Seeing like a state: How certain schemes to improve the human condition have failed*. yale university Press, 2020.
- [369] Nick Seaver. "Studying up: The ethnography of technologists". In: *Ethnography Matters* 10 (2014).
- [370] Treasury Board Secretariat. "Directive on automated decision-making". In: *Ottawa (ON): Government of Canada (modified 2019-02-05)* (2020).
- [371] Neil Selwyn. *Education and Technology: Key Issues and Debates*. Bloomsbury Publishing, 2016.
- [372] Mark P Sendak, William Ratliff, Dina Sarro, Elizabeth Alderton, Joseph Futoma, Michael Gao, Marshall Nichols, Mike Revoir, Faraz Yashar, Corinne Miller, et al. "Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study". In: *JMIR medical informatics* 8.7 (2020), e15182.
- [373] Stefan A Seuring, Julia Koplin, Torsten Behrens, and Uwe Schneidewind. "Sustainability assessment in the German detergent industry: from stakeholder involvement to sustainability indicators". In: *Sustainable Development* 11.4 (2003), pp. 199–212.

- [374] Shreya Shankar, JD Zamfirescu-Pereira, Björn Hartmann, Aditya Parameswaran, and Ian Arawjo. "Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences". In: *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 2024, pp. 1–14.
- [375] Aaron Shapiro. "Reform predictive policing". In: *Nature* 541.7638 (2017), pp. 458–460.
- [376] Amanda Sharkey and Natalie Wood. "The Paro seal robot: Demeaning or enabling". In: *AISB*. 2014.
- [377] Ramzi Shawahna. "Quality indicators of Pharmaceutical Care in Palestinian Integrative Healthcare Facilities: findings of a qualitative study among stakeholders". In: *Evidence-Based Complementary and Alternative Medicine* 2020.1 (2020), p. 4520769.
- [378] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. "Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors". In: *arXiv preprint arXiv:2105.02980* (2021).
- [379] Hua Shen and Ting-Hao'Kenneth' Huang. "Explaining the Road Not Taken". In: *arXiv preprint arXiv:2103.14973* (2021).
- [380] Katie Shilton. "Values levers: Building ethics into design". In: *Science, Technology, & Human Values* 38.3 (2013), pp. 374–397.
- [381] Takuma Shiozawa, Sosei Yamaguchi, Makoto Ogawa, Makiko Abe, Takayuki Kawaguchi, Momoka Igarashi, Naonori Yasuma, and Chiyo Fujii. "Consensus development of priority outcome domains for community mental health cares by multiple stakeholders: Protocol for an online Delphi study in Japan". In: *Neuropsychopharmacology Reports* 41.4 (2021), pp. 554–561.
- [382] Åsa Skjutar, Kyllike Christensson, and Maria Müllersdorf. "Exploring indicators for pain rehabilitation: a Delphi study using a multidisciplinary expert panel". In: *Musculoskeletal Care* 7.4 (2009), pp. 227–242.
- [383] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. "Participation is not a design fix for machine learning". In: *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 2022, pp. 1–6.
- [384] Christopher T Small, Ivan Vendrov, Esin Durmus, Hadjar Homaei, Elizabeth Barry, Julien Cornebise, Ted Suzman, Deep Ganguli, and Colin Megill. "Opportunities and risks of LLMs for scalable deliberation with Polis". In: *arXiv preprint arXiv:2306.11932* (2023).
- [385] C. Estelle Smith, Bowen Yu, Anjali Srivastava, Aaron Halfaker, Loren Terveen, and Haiyi Zhu. "Keeping Community in the Loop: Understanding Wikipedia Stakeholder Values for Machine Learning-Based Systems". In: *Conference on Human Factors in Computing Systems - Proceedings* (2020), pp. 1–14. DOI: [10.1145/3313831.3376783](https://doi.org/10.1145/3313831.3376783).
- [386] Freya Smith, Malak Sadek, and Céline Mougenot. "Empowering end-users in co-designing AI: An AI literacy card-based toolkit for non-technical audiences". In: *36th International BCS Human-Computer Interaction Conference*. BCS Learning & Development. 2023, pp. 13–22.

- [387] Jessie J Smith, Aishwarya Satwani, Robin Burke, and Casey Fiesler. "Recommend me? designing fairness metrics with providers". In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2024, pp. 2389–2399.
- [388] Kaitlin Smith. *Teaching in the Light of Women's History*. 2022. URL: <https://www.facinghistory.org/ideas-week/teaching-light-womens-history>.
- [389] Franchesca Spektor, Sarah E Fox, Ezra Awumey, Ben Begleiter, Chinmay Kulkarri, Betsy Stringam, Christine A Riordan, Hye Jin Rho, Hunter Akridge, and Jodi Forlizzi. "Charting the Automation of Hospitality: An Interdisciplinary Literature Review Examining the Evolution of Frontline Service work in the Face of Algorithmic Management". In: *Proceedings of the ACM on Human-Computer Interaction* 7.CSCW1 (2023), pp. 1–20.
- [390] Katta Spiel, Emeline Brulé, Christopher Frauenberger, Gilles Bailley, and Geraldine Fitzpatrick. "In the details: the micro-ethics of negotiations and in-situ judgements in participatory design with marginalised children". In: *CoDesign* 16.1 (2020), pp. 45–65.
- [391] Clay Spinuzzi. "A Scandinavian challenge, a US response: Methodological assumptions in Scandinavian and US prototyping approaches". In: *ACM SIG-DOC*. 2002.
- [392] Megha Srivastava, Hoda Heidari, and Andreas Krause. "Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019, pp. 2459–2468.
- [393] Barbara White Stack. "CYF: An agency that works, helping kids and their families". In: *Pittsburgh Post-Gazette* (). Online; accessed 8-September-2021. URL: <https://www.post-gazette.com/opinion/Op-Ed/2018/02/11/CYF-An-agency-that-works-helping-kids-and-their-families/stories/201802040037>.
- [394] Eliane Stampfer, Yanjin Long, Vincent Aleven, and Kenneth R Koedinger. "Eliciting intelligent novice behaviors with grounded feedback in a fraction addition tutor". In: *Artificial Intelligence in Education: 15th International Conference, AIED 2011, Auckland, New Zealand, June 28–July 2011* 15. Springer. 2011, pp. 560–562.
- [395] Logan Stapleton, Min Hun Lee, Diana Qing, Marya Wright, Alexandra Chouldechova, Ken Holstein, Zhiwei Steven Wu, and Haiyi Zhu. "Imagining new futures beyond predictive systems in child welfare: A qualitative study with impacted stakeholders". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022, pp. 1162–1177.
- [396] Susan Leigh Star. "The ethnography of infrastructure". In: *ABS* (1999).
- [397] Susan Leigh Star and Anselm Strauss. "Layers of silence, arenas of voice: The ecology of visible and invisible work". In: *CSCW'99* (1999).
- [398] Elizabeth D Steiner and Ashley Woo. "Job-related stress threatens the teacher supply". In: *Rand Corporation* 25 (2021), pp. 08–1.
- [399] David Steinhart. *Juvenile Detention Risk Assessment: A Practice Guide to Juvenile Detention Reform*. 2006. URL: <https://assets.aecf.org/m/resourcedoc/aecf-juveniledetentionriskassessment1-2006.pdf>.

- [400] Megan Stevenson. "Assessing Risk Assessment in Action". In: *Law, Economics, and Business Fellows' Discussion Paper Series 85* 103 (2018), pp. 303–384. URL: <https://scholarship.law.umn.edu/mlr/58/>.
- [401] Megan Stevenson and Jennifer Doleac. "Algorithmic Risk Assessment in the Hands of Humans". In: (2021). URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3489440.
- [402] Jonathan Stray, Alon Halevy, Parisa Assar, Dylan Hadfield-Menell, Craig Boutilier, Amar Ashar, Chloe Bakalar, Lex Beattie, Michael Ekstrand, Claire Leibowicz, et al. "Building human values into recommender systems: An interdisciplinary synthesis". In: *ACM Transactions on Recommender Systems* 2.3 (2024), pp. 1–57.
- [403] Ieva Stupans, Leigh McKaige, and Susanne Owen. "Indicators of a quality clinical placement in pharmacy: Stakeholder perspectives". In: *Journal of Pharmacy Practice and Research* 41.2 (2011), pp. 118–121.
- [404] Hariharan Subramonyam, Colleen Seifert, and Eytan Adar. "Towards a process model for co-creating AI experiences". In: *arXiv preprint arXiv:2104.07595* (2021).
- [405] Lucille Alice Suchman. *Plans and situated actions: The problem of human-machine communication*. Cambridge university press, 1987.
- [406] Lucy Suchman. "Anthropological relocations and the limits of design". In: *Annual review of anthropology* 40.1 (2011), pp. 1–18.
- [407] Lucy Suchman. "Do categories have politics? The language/action perspective reconsidered". In: *CSCW'93* (1993).
- [408] Lucy Suchman. "Making work visible". In: *CACM* (1995).
- [409] Sangho Suh, Michael Lai, Kevin Pu, Steven P Dow, and Tovi Grossman. "StoryEnsemble: Enabling Dynamic Exploration & Iteration in the Design Process with AI and Forward-Backward Propagation". In: *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology* (2025).
- [410] Harini Suresh, Rajiv Movva, Amelia Lee Dogan, Rahul Bhargava, Isadora Cruxên, Ángeles Martínez Cuba, Guilia Taurino, Wonyoung So, and Catherine D'Ignazio. "Towards intersectional feminist and participatory ML: A case study in supporting Femicide Counterdata Collection". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022, pp. 667–678.
- [411] Harini Suresh, Emily Tseng, Meg Young, Mary Gray, Emma Pierson, and Karen Levy. "Participation in the age of foundation models". In: *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2024, pp. 1609–1621.
- [412] Annalisa Szymanski, Simret Araya Gebreegziabher, Oghenemaro Anuyah, Ronald A Metoyer, and Toby Jia-Jun Li. "Comparing criteria development across domain experts, lay users, and models in large language model evaluation". In: *arXiv preprint arXiv:2410.02054* (2024).
- [413] Annalisa Szymanski, Noah Ziemis, Heather A Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A Metoyer. "Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks". In: *Proceedings of the 30th International Conference on Intelligent User Interfaces*. 2025, pp. 952–966.

- [414] Soheyla Taie and Laurie Lewis. "Characteristics of 2020-21 Public and Private K-12 School Teachers in the United States: Results from the National Teacher and Principal Survey." In: *NCES* (2022).
- [415] Mei Tan, Hansol Lee, Dakuo Wang, and Hari Subramonyam. "Is a Seat at the Table Enough? Engaging Teachers and Students in Dataset Specification for ML in Education". In: *Proceedings of the ACM on Human-Computer Interaction* 8.CSCW1 (2024), pp. 1–32.
- [416] Sarah Tan, Julius Adebayo, and Kori Inkpen. "Investigating human + machine complementarity for recidivism predictions". In: *arXiv* (2018). arXiv: [arXiv:1808.09123v2](https://arxiv.org/abs/1808.09123v2).
- [417] Sarah Tan, Julius Adebayo, Kori Inkpen, and Ece Kamar. "Investigating human+ machine complementarity for recidivism predictions". In: *arXiv preprint arXiv:1808.09123* (2018).
- [418] *Teacher Education: From Revolution to Evolution*. 2018. URL: <https://teachereducation.steinhardt.nyu.edu/teacher-training-evolution/>.
- [419] MindShare Technology. *Improving Outcomes using data you already have*. <https://mindshare-technology.com/analytics/>. Online; accessed 8-September-2021.
- [420] Alexandra To, Angela DR Smith, Dilruba Showkat, Adinawa Adjagbodjou, and Christina Harrington. "Flourishing in the everyday: Moving beyond damage-centered design in HCI for BIPOC communities". In: *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 2023, pp. 917–933.
- [421] Songül Tolan, Marius Miron, Emilia Gómez, and Carlos Castillo. "Why machine learning may lead to unfairness: Evidence from risk assessment for juvenile justice in catalonia". In: *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*. 2019, pp. 83–92.
- [422] Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. "What clinicians want: contextualizing explainable machine learning for clinical end use". In: *Machine learning for healthcare conference*. PMLR. 2019, pp. 359–380.
- [423] Philipp Tschandl, Noel Codella, Bengü Nisa Akay, Giuseppe Argenziano, Ralph P Braun, Horacio Cabo, David Gutman, Allan Halpern, Brian Helba, Rainer Hofmann-Wellenhof, et al. "Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study". In: *The Lancet Oncology* 20.7 (2019), pp. 938–947.
- [424] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, John Paoli, Susana Puig, Cliff Rosendahl, H. Peter Soyer, Iris Zalaudek, and Harald Kittler. "Human–computer collaboration for skin cancer recognition". In: *Nature Medicine* 26.8 (2020), pp. 1229–1234. ISSN: 1546170X. DOI: [10.1038/s41591-020-0942-0](https://doi.org/10.1038/s41591-020-0942-0). URL: <http://dx.doi.org/10.1038/s41591-020-0942-0>.
- [425] Emily Tseng, Fabian Okeke, Madeline Sterling, and Nicola Dell. "'We can learn. Why not?' Designing Technologies to Engender Equity for Home Health Aides". In: *CHI'20*. 2020.
- [426] Mark Turner, David Budgen, and Pearl Brereton. "Turning software into a service". In: *Computer* 36.10 (2003), pp. 38–44.

- [427] Marie Utterberg Modén, Martin Tallvid, Johan Lundin, and Berner Lindström. "Intelligent tutoring systems: Why teachers abandoned a technology aimed at automating teaching processes". In: (2021).
- [428] Indrė Žliobaitė. "On the relation between accuracy and fairness in binary classification". In: *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FATML)*. 2015.
- [429] Rhema Vaithianathan, Diana Benavides-Prado, Erin Dalton, Alex Chouldechova, and Emily Putnam-Hornstein. "Using a machine learning tool to support high-stakes decisions in child protection". In: *AI Magazine* 42.1 (2021), pp. 53–60.
- [430] Rhema Vaithianathan, Haley Dinh, Allon Kalisher, Chamari Kithulgoda, Emily Kulick, Megh Mayur, Athena Ning, and Diana Benavides Prado. *Implementing a Child Welfare Decision Aide in Douglas County | Methodology Report*. https://csda.aut.ac.nz/_data/assets/pdf_file/0009/347715/Douglas-County-Methodology_Final_3_02_2020.pdf. Online; accessed 8-September-2021.
- [431] Rhema Vaithianathan, Nan Jiang, Tim Maloney, Parma Nand, and Emily Putnam-Hornstein. *Developing Predictive Risk Models to Support Child Maltreatment Hotline Screening Decisions*. 2017. URL: <https://www.alleghenycountyanalytics.us/wp-content/uploads/2017/04/Developing-Predictive-Risk-Models-package-with-cover-1-to-post-1.pdf>.
- [432] Rhema Vaithianathan, Emily Kulick, Emily Putnam-Hornstein, and Diana Benavides Prado. *Allegheny Family Screening Tool: Methodology, Version 2*. https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/Methodology-V2-from-16-ACDHS-26-PredictiveRisk_Package_050119_FINAL-7.pdf. Online; accessed 8-September-2021.
- [433] Rhema Vaithianathan, Emily Kulick, Emily Putnam-Hornstein, and Diana Benavides Prado. *Allegheny Family Screening Tool: Methodology, Version 2*. https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/Methodology-V2-from-16-ACDHS-26-PredictiveRisk_Package_050119_FINAL-7.pdf. Online; accessed 19-February-2022. 2019.
- [434] Rhema Vaithianathan, Tim Maloney, Emily Putnam-Hornstein, and Nan Jiang. "Children in the public benefit system at risk of maltreatment: Identification via predictive modeling". In: *American journal of preventive medicine* 45.3 (2013), pp. 354–359.
- [435] Rhema Vaithianathan, Emily Putnam-Hornstein, Alexandra Chouldechova, Diana Benavides-Prado, and Rachel Berger. "Hospital injury encounters of children identified by a predictive risk model for screening child maltreatment referrals: evidence from the Allegheny Family Screening Tool". In: *JAMA pediatrics* 174.11 (2020).
- [436] Rhema Vaithianathan, Emily Putnam-Hornstein, Nan Jiang, Parma Nand, and Tim Maloney. "Developing predictive models to support child maltreatment hotline screening decisions: Allegheny County methodology and implementation". In: *Center for Social data Analytics* (2017).
- [437] MW Van Someren, YF Barnard, and JAC Sandberg. "The think aloud method: a practical approach to modelling cognitive". In: *London: AcademicPress* (1994).

- [438] Lisa M Vaughn and Farrah Jacquez. "Participatory research methods—choice points in the research process". In: *Journal of participatory research methods* 1.1 (2020).
- [439] Michael Veale, Max Van Kleek, and Reuben Binns. "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*. 2018, pp. 1–14.
- [440] Harsh Vishwakarma, Ankush Agarwal, Ojas Patil, Chaitanya Devaguptapu, and Mahesh Chandran. "Can LLMs Help You at Work? A Sandbox for Evaluating LLM Agents in Enterprise Environments". In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 2025, pp. 9178–9212.
- [441] Hanna Wallach, Meera Desai, A Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P Alex Dow, et al. "Position: Evaluating generative ai systems is a social science measurement challenge". In: *arXiv preprint arXiv:2502.00561* (2025).
- [442] Hanna Wallach, Meera Desai, Nicholas Pangakis, A Feder Cooper, Angelina Wang, Solon Barocas, Alexandra Chouldechova, Chad Atalla, Su Lin Blodgett, Emily Corvi, et al. "Evaluating Generative AI Systems is a Social Science Measurement Challenge". In: *arXiv preprint arXiv:2411.10939* (2024).
- [443] Angelina Wang, Sayash Kapoor, Solon Barocas, and Arvind Narayanan. "Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy". In: *Available at SSRN* (2022).
- [444] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. "'Brilliant AI Doctor' in Rural Clinics: Challenges in AI-Powered Clinical Decision Support System Deployment". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–18.
- [445] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. "Mmlu-pro: A more robust and challenging multi-task language understanding benchmark". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 95266–95290.
- [446] Zora Zhiruo Wang, Yijia Shao, Omar Shaikh, Daniel Fried, Graham Neubig, and Diyi Yang. "How Do AI Agents Do Human Work? Comparing AI and Human Workflows Across Diverse Occupations". In: *arXiv preprint arXiv:2510.22780* (2025).
- [447] Human Rights Watch. "'Not in it for Justice': How California's Pretrial Detention and Bail System Unfairly Punishes Poor People". In: (2017). URL: <https://www.hrw.org/report/2017/04/11/not-itjustice/how-californias-pretrial-detention-and-bail-system-unfairly>.
- [448] Audrey Watters. *The History of the Future of Education*. 2015.
- [449] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. "The what-if tool: Interactive probing of machine learning models". In: *IEEE transactions on visualization and computer graphics* 26.1 (2019), pp. 56–65.

- [450] Eliane S Wiese and Kenneth R Koedinger. "Designing grounded feedback: Criteria for using linked representations to support learning of abstract symbols". In: *International Journal of Artificial Intelligence in Education* 27 (2017), pp. 448–474.
- [451] Bryan Wilder, Eric Horvitz, and Ece Kamar. "Learning to complement humans". In: *arXiv preprint arXiv:2005.00582* (2020).
- [452] Christine T Wolf, Haiyi Zhu, Julia Bullard, Min Kyung Lee, and Jed R Brubaker. "The changing contours of" participation" in data-driven, algorithmic ecosystems: Challenges, tactics, and an agenda". In: *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 2018, pp. 377–384.
- [453] Richmond Y Wong, Michael A Madaio, and Nick Merrill. "Seeing like a toolkit: How toolkits envision the work of AI ethics". In: *Proceedings of the ACM on Human-Computer Interaction* 7.CSCW1 (2023), pp. 1–27.
- [454] Allison Woodruff, Renee Shelby, Patrick Gage Kelley, Steven Rousso-Schindler, Jamila Smith-Loud, and Lauren Wilcox. "How knowledge workers think generative ai will (not) transform their industries". In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 2024, pp. 1–26.
- [455] James Wright. "Robots vs migrants? Reconfiguring the future of Japanese institutional eldercare". In: *CAS* (2019).
- [456] James Wright. "Tactile care, mechanical Hugs: Japanese caregivers and robotic lifting devices". In: *Asian Anthropology* (2018).
- [457] Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. "Fine-grained human feedback gives better rewards for language model training". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 59008–59033.
- [458] Jeremy C Wyatt and Douglas G Altman. "Commentary: Prognostic models: Clinically useful or quickly forgotten?" In: *BMJ* (1995).
- [459] Qian Yang, Aaron Steinfeld, and John Zimmerman. "Unremarkable AI: Fitting intelligent decision support into critical, clinical decision-making processes". In: *Proceedings of the 2019 CHI conference on human factors in computing systems*. 2019, pp. 1–11.
- [460] Qian Yang, Aaron Steinfeld, and John Zimmerman. "Unremarkable AI: Fitting intelligent decision support into critical, clinical decision-making processes". In: *Conference on Human Factors in Computing Systems - Proceedings* (2019). DOI: [10.1145/3290605.3300468](https://doi.org/10.1145/3290605.3300468). arXiv: [1904.09612](https://arxiv.org/abs/1904.09612).
- [461] Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey, and James F Antaki. "Investigating the heart pump implant decision process: opportunities for decision support tools to help". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2016, pp. 4477–4488.
- [462] Zonghai Yao, Zihao Zhang, Chaolong Tang, Xingyu Bian, Youxia Zhao, Zhichao Yang, Junda Wang, Huixue Zhou, Won Seok Jang, Feiyun Ouyang, et al. "Medqa-cs: Benchmarking large language models clinical skills using an ai-sce framework". In: *arXiv preprint arXiv:2410.01553* (2024).

- [463] Nur Yildirim, Changhoon Oh, Deniz Sayar, Kayla Brand, Supritha Challa, Violet Turri, Nina Crosby Walton, Anna Elise Wong, Jodi Forlizzi, James McCann, et al. "Creating Design Resources to Scaffold the Ideation of AI Concepts". In: *DIS'23* (2023).
- [464] Nur Yildirim, Mahima Pushkarna, Nitesh Goyal, Martin Wattenberg, and Fernanda Viégas. "Investigating How Practitioners Use Human-AI Guidelines: A Case Study on the People+ AI Guidebook". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–13.
- [465] Covenant House New York. *There is No Such Thing As Women's Work: The gender of labor and the devaluation of social workers*. 2022. URL: <https://ny.covenanthouse.org/there-is-no-such-thing-as-womens-work-the-gender-of-labor-and-the-devaluation-of-social-workers/>.
- [466] Meg Young, Upol Ehsan, Ranjit Singh, Emnet Tafesse, Michele Gilman, Christina Harrington, and Jacob Metcalf. "Participation versus scale: Tensions in the practical demands on participatory AI". In: *First Monday* (2024).
- [467] Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. "Artificial intelligence in healthcare". In: *Nature biomedical engineering* 2.10 (2018), pp. 719–731.
- [468] Angie Zhang, Rocita Rana, Alexander Boltz, Veena Dubal, and Min Kyung Lee. "Data Probes as Boundary Objects for Technology Policy Design: Demystifying Technology for Policymakers and Aligning Stakeholder Objectives in Rideshare Gig Work". In: *CHI'24*. 2024.
- [469] Angie Zhang, Olympia Walker, Kaci Nguyen, Jiajun Dai, Anqing Chen, and Min Kyung Lee. "Deliberating with AI: Improving Decision-Making for the Future through Participatory AI Design and Stakeholder Deliberation". In: *Proceedings of the ACM on Human-Computer Interaction* 7.CSCW1 (2023), pp. 1–32.
- [470] Qiaoning Zhang, Matthew L Lee, and Scott Carter. "You Complete Me: Human-AI Teams and Complementary Expertise". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–28.
- [471] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. "Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 295–305.
- [472] Yunfeng Zhang, Q. Vera Liao, and Rachel K.E. Bellamy. "Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making". In: *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020), pp. 295–305. DOI: [10.1145/3351095.3372852](https://doi.org/10.1145/3351095.3372852). arXiv: [2001.02114](https://arxiv.org/abs/2001.02114).
- [473] Dora Zhao, Jerone TA Andrews, Orestis Papakyriakopoulos, and Alice Xiang. "Position: Measure Dataset Diversity, Don't Just Claim It". In: *arXiv preprint arXiv:2407.08188* (2024).
- [474] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. "Value-sensitive algorithm design: Method, case study, and lessons". In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (2018), pp. 1–23.
- [475] Alexandra Zyttek, Dongyu Liu, Rhema Vaithianathan, and Kalyan Veeramachaneni. "Sibyl: Understanding and addressing the usability challenges of machine learning in high-stakes decision making". In: *IEEE Transactions on Visualization and Computer Graphics* (2021).