"Why Do I Care What's Similar?" Probing Challenges in AI-Assisted Child Welfare Decision-Making through Worker-AI Interface Design Concepts

Anna Kawakami*
Carnegie Mellon University
Pittsburgh, USA
akawakam@andrew.cmu.edu

Venkatesh Sivaraman*
Carnegie Mellon University
Pittsburgh, USA
vsivaram@andrew.cmu.edu

Logan Stapleton
University of Minnesota
Minneapolis, USA
stapl158@umn.edu

Hao-Fei Cheng
Carnegie Mellon University
Pittsburgh, USA
haofeic@andrew.cmu.edu

Adam Perer
Carnegie Mellon University
Pittsburgh, USA
adamperer@cmu.edu

Zhiwei Steven Wu
Carnegie Mellon University
Pittsburgh, USA
zstevenwu@cmu.edu

Haiyi Zhu
Carnegie Mellon University
Pittsburgh, USA
haiyiz@cs.cmu.edu

Kenneth Holstein
Carnegie Mellon University
Pittsburgh, USA
kjholste@cs.cmu.edu

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.

CCS CONCEPTS

• **Human-centered computing** → **Human Computer Interaction (HCI)**; **Interactive system and tools**.

*Both co-first authors contributed equally to this research.

This work is licensed under a Creative Commons Attribution International 4.0 License.

DIS '22, June 13-June 17, 2022,
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9358-4/22/06.
https://doi.org/10.1145/3532106.3533556

KEYWORDS

AI-assisted decision-making, algorithmic decision support, design, child welfare, social work

ACM Reference Format:
Anna Kawakami, Venkatesh Sivaraman, Logan Stapleton, Hao-Fei Cheng, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. "Why Do I Care What's Similar?" Probing Challenges in AI-Assisted Child Welfare Decision-Making through Worker-AI Interface Design Concepts. In *DIS Conference on Designing Interactive Systems (DIS '22), June 13-June 17, 2022.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3532106.3533556

1 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 [3, 37]. AI-based tools are commonly introduced into social decision-making settings with the promise of overcoming human limitations and biases [34, 58]. However, in such settings, AI judgments are themselves likely to be imperfect and biased, even if in different ways than humans [1, 17, 24]. Prior work across a range of social decision-making contexts indicates potential for frontline decision-makers and ADS tools to complement each other's capabilities and help to overcome each other's limitations (e.g., [8, 14, 21, 31]). 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., [20, 44, 51, 52]).

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 [14, 32, 40]. 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 [19] and child maltreatment screening [14]. As a result, the growing body of AI design knowledge obtained from less contentious and value-laden decision settings [35] 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 [10, 57]. As we discuss in [32], 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 [32], 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 BACKGROUND AND RELATED WORK

2.1 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., [10, 23, 43, 52]). 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 [2, 14, 21]. 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., [4, 15, 20, 36]).

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., [8, 14, 21]). 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., [2, 44]). By contrast, some recent work signals potential for relatively simple cognitive cues or reflection prompts to help foster more effective use of ADS [5, 23, 44]. It remains an open research question how best to foster human–algorithm synergy, particularly in real-world AI-assisted decision-making contexts.

2.2 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 [37, 40], 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 [48]), 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 [42, 54]. 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 [19, 34] or represent an imperfect proxy for what decision-makers actually predict [18, 28]. 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 [39, 40, 45]. 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 [32, 48].

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 [10, 46, 47, 60]. 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. [60] 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.

3 STUDY CONTEXT

3.1 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" [10, 49]. 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 1, 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) [57]. In Figure 2, we overview how the AFST is 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

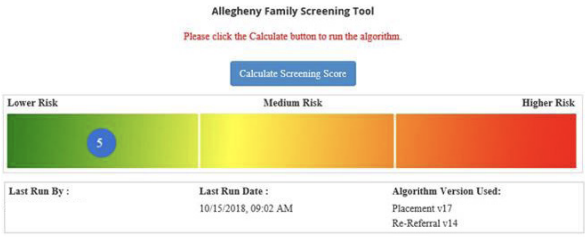

Figure 1: A screenshot showing the AFST interface that the call screeners and supervisors see when using the tool [56]. The AFST risk score outputs a single number between 1 (low risk of future placement, left-most side of bar) and 20 (high risk of future placement, right-most side of bar) for every case. In the image, the tool gave a risk score of 5.

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.

3.2 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 [9, 10, 14, 55]. Other public sector agencies are beginning to view the AFST as an example of what ADS in social work could or should look like [46]. However, most existing studies on the impact of the AFST on workers' screening decisions are based on analyses of historical, retrospective administrative data [8, 14]. 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 [32], 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 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 [32] inspired specific design concepts in Section 5.

4 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 [32]. 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. [12, 24]). 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 5 and visually summarized in Figure 10.

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 1 shows participant demographics, presented in aggregate form to avoid making individual workers identifiable within their workplace.

4.1 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 [32]. Following the first visit, our research team conducted Interpretation Sessions [26] 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 [16, 23]. 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.

4.2 Analysis

Following [16, 24], we analyzed participants' feedback on each design concept separately, while also examining trends in their overall feedback across concepts and iterations (see Figure 10). 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.

4.3 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.

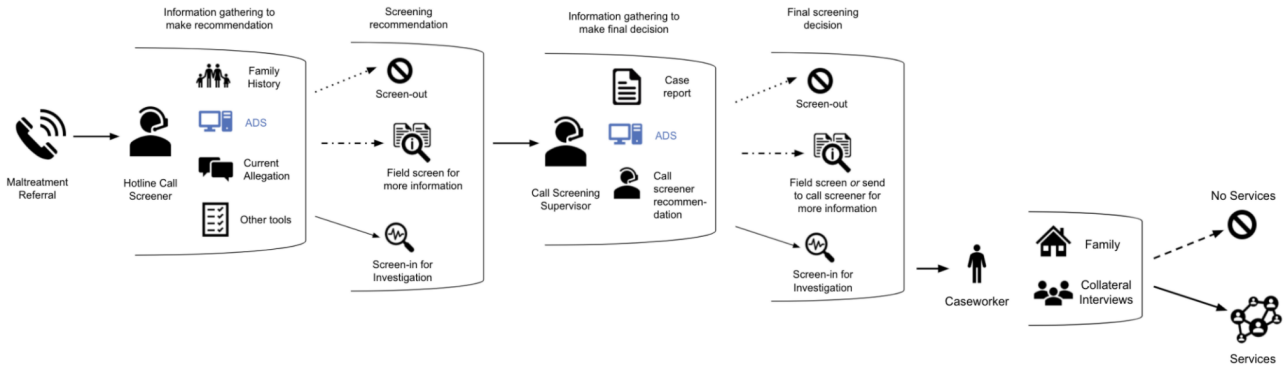

Figure 2: Diagram showing a high-level overview of Allegheny County's child maltreatment screening process, illustrating when an ADS assists call screeners' and supervisors' screening decisions. Taken from [32].

4.4 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.

5 DESIGN CONCEPTS

In this section, we briefly introduce our ten design concepts, seven of which had visual representations (see Figures 3-9). 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**.

5.1 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 [32]. Therefore, we explored three design concepts that surface discrepancies between social workers' and the AI's decision processes [14, 22].

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 3). 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 worker's recommendation is at odds with the AFST score (depicted in Figure 4). 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 [13]). 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.

5.2 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 [32]. 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.

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

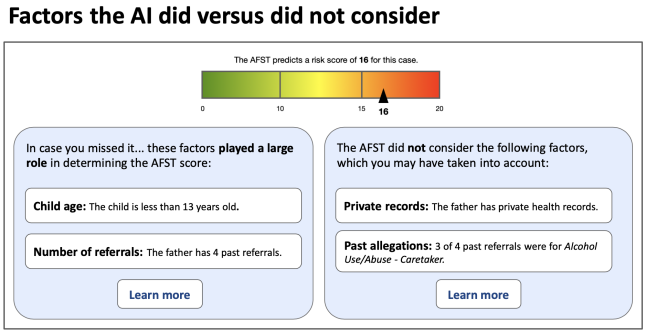

Figure 3: This design concept invited workers to consider not only factors that contributed to an AFST prediction, but also factors that the AFST is unable to take into account.

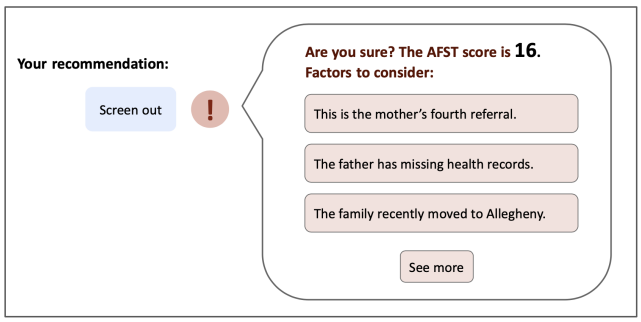

Figure 4: In this design concept, the worker would be prompted to consider additional factors if their assessment is significantly different from the algorithm's.

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?'"* [32]. 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 5), and were asked to imagine how they might use it to inform their screening decisions.

5.3 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., [7, 29, 53]. 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 [14, 19]. 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.

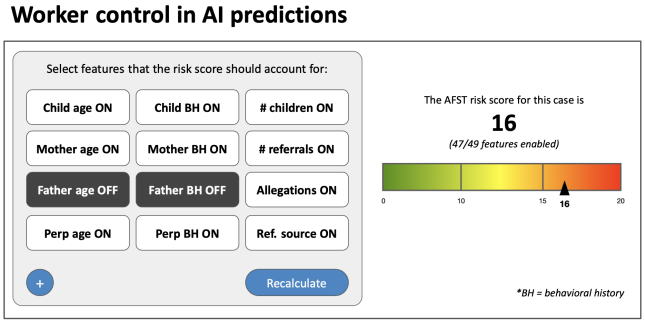

Figure 5: This design concept, inspired by worker feedback in the first round, would allow workers to dynamically adjust the factors or individuals that the AFST uses when computing a risk score.

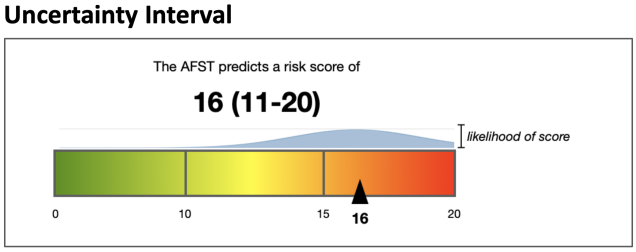

Figure 6: This design concept draws workers' attention to the inherent variability and uncertainty in outcomes predicted by the AFST, using a conventional probability distribution superimposed on the score.

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 6). Although many more effective uncertainty visualization techniques have been developed [27, 30], 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 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 7), 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 [44].

5.4 Measurement and feedback on historical decisions

As described in [32], workers described that their existing performance measures caused them to feel organizational pressures to avoid disagreeing with the AFST score "too often," incentivizing

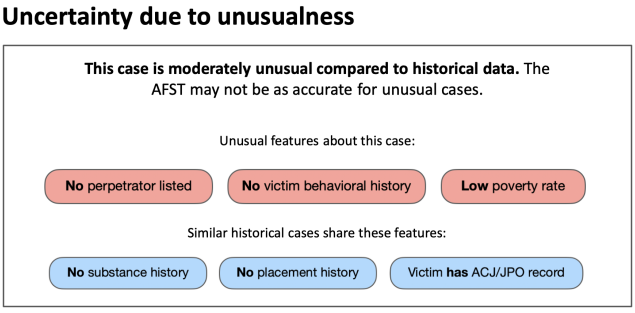

Figure 7: In this design concept, we explore notions of epistemic uncertainty related to the unusualness of a case, and provide feature explanations of why a case could be considered unusual.

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 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 8. 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 [11, 13, 14] (see Figure 9).

6 RESULTS

In this section, we present findings organized along our four design topics. Figure 10 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

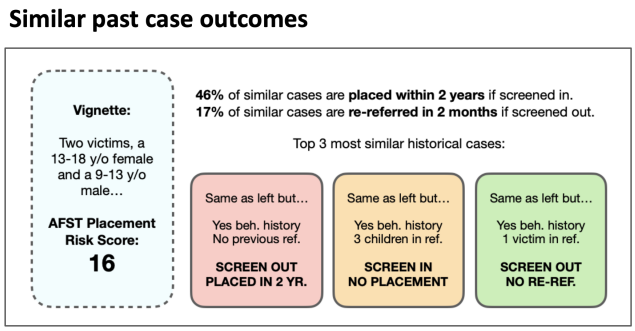

Figure 8: This design concept portrays AFST scores as founded on patterns in historical data using observed outcomes from similar cases, including the top three 'nearest neighbors' of the current case.

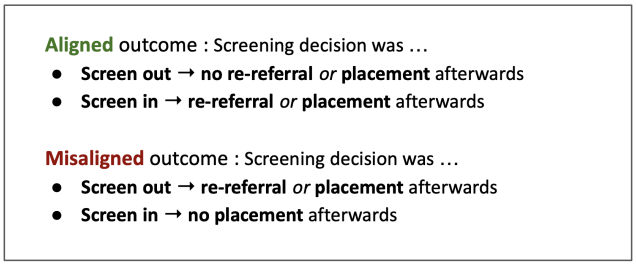

Figure 9: This design concept maps out a possible way to measure whether a worker's decision is aligned or misaligned with the final outcome, incorporating commonly-used placement and re-referral metrics.

of the overall concept (border color) did not match their assessment of the specific design sketch presented (fill color). As described in Section 4, because we conducted the design interviews during participants' working hours, some were unable to complete all study activities.

6.1 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.

6.1.1 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.

6.1.2 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. [6, 23, 38]). 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*

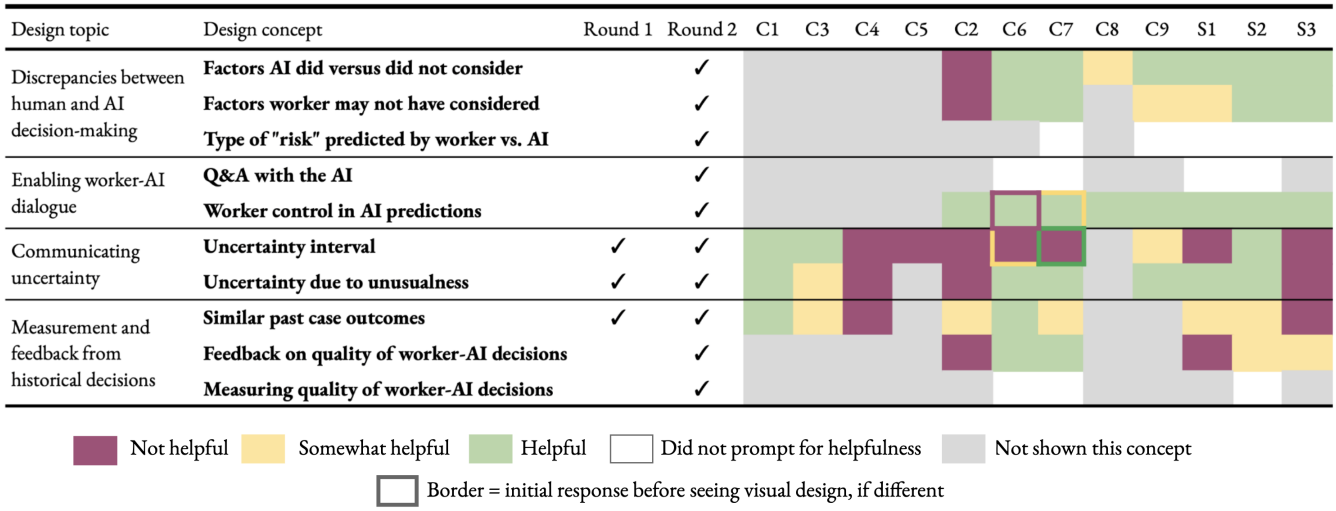

Figure 10: Matrix showing overall perceived helpfulness for each participant (columns) and design concept (rows). As shown by the checkmarks, three design concepts were introduced during the first visit's design interviews. The remaining seven design concepts were introduced during the second visit's design interviews, informed by findings from the field observations and interviews from the first visit.

> *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."*

6.1.3 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."*

6.2 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.

6.2.1 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. [38]). 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."*

6.2.2 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.

6.3 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.

6.3.1 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."*

6.3.2 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.

6.4 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.

6.4.1 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).

6.4.2 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 [32]. 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.

6.4.3 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 *"harshest 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.

7 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 6.1), potential ways to engage in dialogue with an ADS (Section 6.2), various notions of uncertainty (Section 6.3), and measurement and feedback on historical AI-assisted worker decisions (Section 6.4). 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.

7.1 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).

7.1.1 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 [50]. 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 [25], 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.*

7.1.2 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 [38, 50]. 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 [48]. 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.*

7.1.3 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., [30]). Data visualization research has often focused on this conventional depiction of uncertainty, improving its efficacy through the use of discrete representations and animations [27, 30, 33]. 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 [41]. 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 [2, 59], 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 [7, 29, 53]. 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.*

7.1.4 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 [9], 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. [11] 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'a 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.*

7.2 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 [60], 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 [48] 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 [18, 52]. 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 [57]), 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 [53, 59], 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 [8, 14], 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.

7.3 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.

8 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.

ACKNOWLEDGMENTS

This work was supported by National Science Foundation (NSF) under Award No. 1939606, 2001851, 2000782 and 1952085 and the Carnegie Mellon University Block Center for Technology and Society Award No. 53680.1.5007718. This work would not be possible without the support and feedback provided by our collaborators at Allegheny County's Department of Human Services, including all of the call screeners and supervisors interviewed for this study.

REFERENCES

[1] Ali Alkhatib and Michael Bernstein. 2019. 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*. 1–13.

[2] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. 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* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. https://doi.org/10.1145/3411764.3445717

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

[4] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th international conference on intelligent user interfaces*. 454–464.

[5] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. 5, April (2021). https://doi.org/10.1145/3449287 arXiv:2102.09692

[6] Susan Bull and Judy Kay. 2016. SMILI©: A framework for interfaces to learning data in open learner models, learning analytics and related fields. *International Journal of Artificial Intelligence in Education* 26, 1 (2016), 293–331. https://doi.org/10.1007/s40593-015-0090-8

[7] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[8] Hao-Fei Cheng, Logan Stapleton, Anna Kawakami, Venkatesh Sivaraman, Yanghuidi Cheng, Diana Qing, Adam Perer, Ken Holstein, Zhiwei Steven Steven Wu, and Haiyi Zhu. 2022. How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022), 1–17.

[9] Hao-Fei Cheng, Logan Stapleton, Ruiqi Wang, Paige Bullock, Alexandra Chouldechova, Zhiwei Steven Steven Wu, and Haiyi Zhu. 2021. Soliciting Stakeholders' Fairness Notions in Child Maltreatment Predictive Systems. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), 1–17. https://doi.org/10.1145/3411764.3445308

[10] Alexandra Chouldechova, Emily Putnam-Hornstein, Suzanne Dworak-Peck, Diana Benavides-Prado, Oleksandr Fialko, Rhema Vaithianathan, Sorelle A Friedler, and Christo Wilson. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *Proceedings of Machine Learning Research* 81 (2018), 1–15. http://proceedings.mlr.press/v81/chouldechova18a.html

[11] Amanda Coston, Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. 2020. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 582–593.

[12] Scott Davidoff, Min Kyung Lee, Anind K Dey, and John Zimmerman. 2007. Rapidly exploring application design through speed dating. In *International Conference on Ubiquitous Computing*. Springer, 429–446.

[13] Maria De-Arteaga, Artur Dubrawski, and Alexandra Chouldechova. 2021. Leveraging Expert Consistency to Improve Algorithmic Decision Support. *arXiv preprint arXiv:2101.09648* (2021).

[14] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. 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*. 1–12.

[15] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[16] Tawanna R Dillahunt, Jason Lam, Alex Lu, and Earnest Wheeler. 2018. Designing future employment applications for underserved job seekers: a speed dating study. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 33–44.

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

[18] Riccardo Fogliato, Alice Xiang, Daniel Nagin, and Alexandra Chouldechova. 2021. *On the Validity of Arrest as a Proxy for Offense : Race and the Likelihood of Arrest for Violent Crimes*. Association for Computing Machinery. arXiv:arXiv:2105.04953v1

[19] Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. 2021. 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*. 1–14.

[20] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019). https://doi.org/10.1145/3359152

[21] Kenneth Holstein and Vincent Aleven. 2021. Designing for human-AI complementarity in K-12 education. *arXiv preprint arXiv:2104.01266* (2021).

[22] Kenneth Holstein, Vincent Aleven, and Nikol Rummel. 2020. A Conceptual Framework for Human–AI Hybrid Adaptivity in Education. In *International Conference on Artificial Intelligence in Education*. Springer, 240–254.

[23] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. 2019. Co-designing a real-time classroom orchestration tool to support teacher–AI complementarity. *Journal of Learning Analytics* 6, 2 (2019).

[24] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. 2019. Designing for complementarity: Teacher and student needs for orchestration support in AI-enhanced classrooms. In *International Conference on Artificial Intelligence in Education*. Springer, 157–171.

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

[26] Karen Holtzblatt and Hugh Beyer. 1997. *Contextual design: defining customer-centered systems*. Elsevier.

[27] Jessica Hullman, Paul Resnick, and Eytan Adar. 2015. Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PloS one* 10, 11 (2015), e0142444.

[28] Abigail Z Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 375–385.

[29] Ekaterina Jussupow, Kai Spohrer, Armin Heinzl, and Joshua Gawlitza. 2021. Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. *Information Systems Research* (2021).

[30] Alex Kale, Matthew Kay, and Jessica Hullman. 2020. Visual reasoning strategies for effect size judgments and decisions. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 272–282.

[31] Ece Kamar. 2016. Directions in hybrid intelligence: Complementing AI systems with human intelligence. *International Joint Conference on Artificial Intelligence (IJCAI)* (2016), 4070–4073.

[32] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. arXiv preprint arXiv:2204.02310* (2022).

[33] Matthew Kay, Tara Kola, Jessica Hullman, and Sean A. Munson. 2016. 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*. 5092–5103. https://doi.org/10.1145/2858036.2858558

[34] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.

[35] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies. *arXiv preprint arXiv:2112.11471* (2021).

[36] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[37] Victoria Lee and Lasana Harris. 2013. How social cognition can inform social decision making. *Frontiers in neuroscience* 7 (2013), 259.

[38] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[39] Kristian Lum, David B. Dunson, and James Johndrow. 2021. Closer than they appear: A Bayesian perspective on individual-level heterogeneity in risk assessment. *arXiv* (2021). arXiv:2102.01135 http://arxiv.org/abs/2102.01135

[40] Arvind Narayanan. 2019. How to recognize AI snake oil. *Arthur Miller Lecture on Science and Ethics* (2019).

[41] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, Sebastian Nowozin, Joshua V Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. *Advances in Neural Information Processing Systems* NeurIPS (2019).

[42] Bhavik N Patel, Louis Rosenberg, Gregg Willcox, David Baltaxe, Mimi Lyons, Jeremy Irvin, Pranav Rajpurkar, Timothy Amrhein, Rajan Gupta, Safwan Halabi, et al. 2019. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ digital medicine* 2, 1 (2019), 1–10.

[43] 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. 2019. Human-machine partnership with artificial intelligence for chest radiograph diagnosis. *npj Digital Medicine* (2019). https://doi.org/10.1038/s41746-019-0189-7

[44] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–52.

[45] Matthew J. Salganik, Ian Lundberg, Alexander T. Kindel, Caitlin E. Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M. Altschul, Jennie E. Brand, Nicole Bohme Carnegie, Ryan James Compton, Debanjan Datta, Thomas Davidson, Anna Filippova, Connor Gilroy, Brian J. Goode, Eaman Jahani, Ridhi Kashyap, Antje Kirchner, Stephen McKay, Allison C. Morgan, Alex Pentland, Kivan Polimis, Louis Raes, Daniel E. Rigobon, Claudia V. Roberts, Diana M. Stanescu, Yoshihiko Suhara, Adaner Usmani, Erik H. Wang, Muna Adem, Abdulla Alhajri, Bedoor AlShebli, Redwane Amin, Ryan B. Amos, Lisa P. Argyle, Livia Baer-Bositis, Moritz Büchi, Bo-Ryehn Chung, William Eggert, Gregory Faletto, Zhilin Fan, Jeremy Freese, Tejomay Gadgil, Josh Gagné, Yue Gao, Andrew Halpern-Manners, Sonia P. Hashim, Sonia Hausen, Guanhua He, Kimberly Higuera, Bernie Hogan, Ilana M. Horwitz, Lisa M. Hummel, Naman Jain, Kun Jin, David Jurgens, Patrick Kaminski, Areg Karapetyan, E. H. Kim, Ben Leizman, Naijia Liu, Malte Möser, Andrew E. Mack, Mayank Mahajan, Noah Mandell, Helge Marahrens, Diana Mercado-Garcia, Viola Mocz, Katariina Mueller-Gastell, Ahmed Musse, Qiankun Niu, William Nowak, Hamidreza Omidvar, Andrew Or, Karen Ouyang, Katy M. Pinto, Ethan Porter, Kristin E. Porter, Crystal Qian, Tamkinat Rauf, Anahit Sargsyan, Thomas Schaffner, Landon Schnabel, Bryan Schonfeld, Ben Sender, Jonathan D. Tang, Emma Tsurkov, Austin van Loon, Onur Varol, Xiafei Wang, Zhi Wang, Julia Wang, Flora Wang, Samantha Weissman, Kirstie Whitaker, Maria K. Wolters, Wei Lee Woon, James Wu, Catherine Wu, Kengran Yang, Jingwen Yin, Bingyu Zhao, Chenyun Zhu, Jeanne Brooks-Gunn, Barbara E. Engelhardt, Moritz Hardt, Dean Knox, Karen Levy, Arvind Narayanan, Brandon M. Stewart, Duncan J. Watts, and Sara McLanahan. 2020. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences* 117, 15 (2020), 8398–8403. https://doi.org/10.1073/pnas.1915006117 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1915006117

[46] Anjana Samant, Aaron Horowitz, Kath Xu, and Sophie Beiers. 2021. Family surveillance by algorithm: The rapidly spreading tools few have heard of. *American Civil Liberties Union (ACLU)* (2021). https://www.aclu.org/sites/default/files/field_document/2021.09.28a_family_surveillance_by_algorithm.pdf

[47] Devansh Saxena, Karla Badillo-Urquiola, Pamela J Wisniewski, and Shion Guha. 2020. A Human-Centered Review of Algorithms used within the US Child Welfare System. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

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

[49] China Scherz. 2011. Protecting children, preserving families: Moral conflict and actuarial science in a problem of contemporary governance. *PoLAR: Political and Legal Anthropology Review* 34, 1 (2011), 33–50.

[50] Hua Shen and Ting-Hao'Kenneth' Huang. 2021. Explaining the Road Not Taken. *arXiv preprint arXiv:2103.14973* (2021).

[51] Megan T Stevenson and Jennifer L Doleac. 2021. Algorithmic risk assessment in the hands of humans. *Available at SSRN 3489440* (2021).

[52] Sarah Tan, Julius Adebayo, and Kori Inkpen. 2018. Investigating human + machine complementarity for recidivism predictions. *arXiv* (2018). arXiv:arXiv:1808.09123v2

[53] Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. 2019. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*. PMLR, 359–380.

[54] 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. 2020. Human–computer collaboration for skin cancer recognition. *Nature Medicine* 26, 8 (2020), 1229–1234. https://doi.org/10.1038/s41591-020-0942-0

[55] Rhema Vaithianathan, Diana Benavides-Prado, Erin Dalton, Alex Chouldechova, and Emily Putnam-Hornstein. 2021. Using a machine learning tool to support high-stakes decisions in child protection. *AI Magazine* 42, 1 (2021), 53–60.

[56] Rhema Vaithianathan, Emily Kulick, Emily Putnam-Hornstein, and Diana Benavides Prado. 2019. 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.

[57] Rhema Vaithianathan, Emily Putnam-Hornstein, Nan Jiang, Parma Nand, and Tim Maloney. 2017. Developing predictive models to support child maltreatment hotline screening decisions: Allegheny County methodology and implementation. *Center for Social data Analytics* (2017).

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

[59] Yunfeng Zhang, Q. Vera Liao, and Rachel K.E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020), 295–305. https://doi.org/10.1145/3351095.3372852 arXiv:2001.02114

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