Training Towards *Critical Use*: Learning to Situate AI Predictions Relative to Human Knowledge

Anna Kawakami
Carnegie Mellon University
Pittsburgh, PA, USA
akawakam@cs.cmu.edu

Luke Guerdan
Carnegie Mellon University
Pittsburgh, USA
lguerdan@cs.cmu.edu

Yanghuidi Cheng
Carnegie Mellon University
Pittsburgh, USA
ycheng8610@gmail.com

Matthew Lee
Toyota Research Institute
Los Altos, CA, USA
matt.lee@tri.global

Scott Carter
Toyota Research Institute
Los Altos, CA, USA
scott.carter@tri.global

Nikos Arechiga
Toyota Research Institute
Los Altos, CA, USA
nikos.arechiga@tri.global

Kate Glazko
University of Washington
Seattle, USA
glazko@cs.washington.edu

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

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

ABSTRACT

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

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; **HCI theory, concepts and models**; **User studies**.

KEYWORDS

algorithm-assisted decision-making, augmented intelligence, human-AI complementarity, AI onboarding and training

ACM Reference Format:
Anna Kawakami, Luke Guerdan, Yanghuidi Cheng, Matthew Lee, Scott Carter, Nikos Arechiga, Kate Glazko, Haiyi Zhu, and Kenneth Holstein. 2023. Training Towards *Critical Use*: Learning to Situate AI Predictions Relative to Human Knowledge . In *Collective Intelligence Conference (CI ’23), November 06–09, 2023, Delft, Netherlands.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3582269.3615595

1 INTRODUCTION

AI-based decision support (ADS) tools are increasingly deployed to assist frontline professionals in critical, real-world contexts like social services, healthcare, and criminal justice, with the hope of improving decision quality. However, realizing these tools’ potential in practice is far from guaranteed [4, 15, 20, 24, 38]. To support human decision-makers in effectively using such tools, it is critical that they receive sufficient training before using them in practice. Yet, today, ADS tools are often introduced into these contexts without adequately preparing the frontline professionals who are asked to use them day-to-day [6, 30, 51].

To address this real-world need, a growing body of studies have explored onboarding and training approaches to support human decision-makers in learning to use ADS tools, demonstrating promising initial results [e.g., 1, 43]. Many of these research studies state a motivation to improve human-AI decision quality in critical applications (e.g., healthcare, criminal justice, and social services). However, a careful review of these studies surfaces key differences between their task designs versus the actual AI-assisted decision-making tasks frontline professionals perform in their day-to-day work [21]:

[*] Both co-senior authors contributed equally to this research.

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

CI ’23, November 06–09, 2023, Delft, Netherlands
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0113-9/23/11.
https://doi.org/10.1145/3582269.3615595

- **Target-construct mismatch:** Whereas most experimental studies of AI-assisted decision-making assume that human experts and AI models are predicting the same construct, in practice models are often trained to predict an imperfect proxy for the true outcome of interest to human decision-makers [13, 21, 28, 32, 57]. In these settings, effective use of AI predictions depends critically on humans' ability to account for such misalignment.
- **Information asymmetry:** While experimental studies typically assume that humans and AI models have access to the same information for decision-making, in practice humans frequently have access to *complementary* information [23, 25, 26, 30]. In these settings, effective use of AI predictions depends critically on humans' ability to make use of decision-relevant information to which they have unique access.

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

In this paper, we explore these questions in a complex, real-world domain involving both properties described above: AI-assisted child maltreatment screening. The use of AI-based decision support tools in child welfare is rapidly spreading [50, 52]. Given significant target-construct mismatch and information asymmetry, it is not advisable in this context to train humans to make decisions that perfectly align with the imperfect proxy used by the AI model [8, 14, 30]. For the same reason, it is unclear how to apply existing measures of "appropriate reliance" on AI predictions in this context, given that these typically assume that alignment with model's target is the goal. To address this gap, we define the notion of ***critical use***—that is, humans' ability to situate AI predictions against potentially complementary knowledge uniquely available to them (but not the AI model).

Adopting *critical use* as a lens, we conducted a within-subjects, randomized controlled experiment to investigate *what* and *how* people learn through practice and feedback on decision-making, in the context of AI-assisted child maltreatment screening. We designed training activities that simulate the actual tasks workers face in a real-world deployment context—Allegheny County's use of the Allegheny Family Screening Tool (AFST) [12, 18, 30]—but provide participants with accelerated opportunities to practice making AI-assisted decisions, low-stakes setting.

To inform our study design and analysis, we drew upon an extensive body of findings from prior field studies in the AFST context, examining how experienced workers make AI-assisted decisions in this setting [8, 10, 11, 14, 30]. Prior literature indicates that experienced workers in this context use the AFST's predictions in sophisticated ways, which improve decision-making. For example, quantitative analyses of workers' decision-making over a multi-year period found that, when workers exercised their discretion to disagree with model predictions on a case-by-case basis, this has the aggregate effect of reducing errors and racial disparities in decision-making, compared with what would have happened had they uncritically agreed with the model's predictions [8, 14]. In line with our definition of *critical use*, field observations revealed that workers drew upon their knowledge of information unavailable to the AFST (e.g., relevant context from qualitative case narratives) to calibrate their reliance upon the model's predictions in each case [30]. To support analysis of study participants' learning, we leveraged our knowledge of experienced workers' AI-assisted decision-making practices to define *indicators* of critical use in this domain.

Overall, our analyses suggest that our training activities promoted more critical use of AI predictions. In earlier practice opportunities, participants in both conditions started off relying on AI model predictions more heavily. They did so even on cases for which the model erred with respect to the proxy it was trained on. However, with increased practice, participants learned to disagree with the AI predictions more often. In particular, participants learned to make decisions that were more accurate with respect to the decision-making of actual experienced workers. The training also improved participants' ability to predict the AI model's behavior. Interestingly, participants who received practice alone saw greater improvements in their ability to predict model behavior, compared to participants who also received explicit feedback alongside practice. Moreover, providing explicit feedback did not impact their learning with respect to decision-making, compared with the effects of repeated practice alone. Our analyses indicate that participants used the qualitative case narratives as a powerful form of *implicit* feedback on the reliability of individual AI predictions, through which participants learned to calibrate their reliance on the AI system. This suggest that, beyond exploring mechanisms for *explicit* feedback on human- and AI-based decisions, future work on training for AI-assisted decision-making should explore the design of rich, implicit feedback mechanisms that empower learners to cross-check AI predictions against concrete, qualitative representations [cf. 55]. More broadly, our findings open up a space of new opportunities and open questions for the design of effective training materials and evaluation metrics for AI-assisted decision-making.

This paper makes the following contributions:

- We introduce the **concept of *critical use***, which can be applied to AI-assisted decision-making settings where traditional measures of "appropriate reliance," based on available ground truth labels, are likely to be unreliable. As a learning goal, critical use is applicable to a range of real-world contexts that involve target-construct mismatch and information asymmetry (e.g., healthcare, criminal justice, content moderation, and education). We explore **how critical use can be measured** via a case study in the child maltreatment screening context.
- We present the **first experimental investigation** in the literature of *what* and *how* humans learn through practice and feedback in settings where human knowledge complements that of an AI model (through target-construct mismatch and information asymmetry). Through our analyses, we investigate **how critical use can be supported through training**.
- Based on our findings, we highlight **new opportunities and open questions** for the design of effective training

materials and learning measures in real-world AI-assisted decision-making settings.

2 BACKGROUND AND RESEARCH QUESTIONS

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

2.1 Training for AI-Assisted Decision-Making

Recent years have seen the rapid adoption of AI-based decision support (ADS) tools to assist human decision-making in settings like social services, education, healthcare, and criminal justice [3, 12, 39, 60]. Despite rapid growth in adoption, frontline professionals are often asked to use these tools day-to-day without adequate training. Even in high-stakes contexts, where AI-assisted decisions have the potential to change the trajectory of individuals' lives, the implementation of training approaches has severely lagged behind investments in model development [6, 7, 30, 51]. This lag has also been recognized in recent regulatory and policy efforts, for example, through government directives requiring employee training on AI-based decision-making tools [54].

To address this gap, an emerging line of research in HCI and CSCW has begun to explore the design of onboarding and training interventions to support humans in learning to work effectively with ADS tools. For example, Cai et. al. conducted a qualitative study with pathologists to identify their needs and desires for onboarding to human-AI collaborative decision-making. Other research has begun to investigate the effectiveness of different training approaches via online experiments. For example, Monzannar et. al. introduced an exemplar-based interactive training approach to hone human decision-makers' mental models of a question answering model's strengths and weaknesses, and explored strategies for selecting training examples that could help humans learn an AI model's error boundaries most efficiently [43]. Lai et. al. explored model-based tutorials as an intervention to improve humans' abilities to understand patterns of model behavior, using a case study in deceptive hotel review detections [37]. Overall, this line of work has shown promising early results, suggesting that the proposed training approaches can support human decision-makers in learning to make more accurate AI-assisted decisions.

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

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

While there is an expectation that both human and AI-based judgements are imperfect, there is hope that careful integrations of the two can improve decision-making by drawing upon the *complementary* strengths of each [2, 16, 27, 30, 59, 61]. Prior literature on measurement in AI-assisted decision-making [21, 32, 49] has articulated several properties of real-world decision-making tasks that can support to human-AI complementarity in practice. In this paper we focus closely on two of these::

- **Target-construct mismatch**: In many real-world deployment settings, AI models are trained to predict *imperfect proxies* for the true construct of interest to decision-makers [13, 21, 28, 32, 57]. For example, in healthcare, a widely-deployed ADS tool was trained to predict *medical costs* as a proxy for clinicians' actual decision-making goal of assessing patients' *medical need* [45]. However, an analysis found that medical cost is a systematically worse proxy for medical need among certain demographic groups—for instance, hospitals may have historically turned patients away from care due to a lack of insurance or other discriminatory factors [46].
- **Information asymmetry**: In real-world settings, humans and AI models frequently have access to complementary sources of information [25, 30, 32]. For example, clinicians make decisions about medical resource allocation across high-risk patients by drawing upon a patient's physical presentation, subjective assessments of their well-being, and real-time test results—information that may be unobservable to an AI model [44]. .

Beyond the healthcare examples provided above, these two properties are observed across a wide range of other real-world AI-assisted decision-making settings, including child welfare, criminal justice, online content moderation, education, lending, and hiring, to name a few. Recent field research has shown that, in some settings, frontline professionals are aware of information asymmetries and misalignments between their own goals as human decision-makers versus the proxy outcomes that an AI model is trained to predict. Furthermore, this awareness can shape how they calibrate their reliance on AI predictions on a case-by-case basis [e.g., 8, 24, 30]. Yet to date, most existing experimental research on AI-assisted decision-making has constructed study environments where these properties are absent. In fact, in a recent review of the literature, Guerdan et. al. [21] found that 92% of experimental studies make the assumption that no target-construct mismatch orinformation asymmetries are present.

Prior studies exploring ways to support humans in learning to make AI-assisted decisions involve tasks such as identifying defective objects [1], passage-based question answering [43], comparing nutritional content [19], or identifying deceptive hotel reviews [37]. In each of these decision tasks, there is no information asymmetry: the human decision-maker is only provided access to the same or a subset of the type of information the AI model is trained on. There is also no pronounced target-construct mismatch: the AI model used in the task is trained on a predictive target (e.g., accuracy of AI-assisted decisions about the *nutritional content* of a given meal [19], or whether an *object is defective* [1]) that directly corresponds to both the AI and the human's predictive target. Like other studies evaluating AI-assisted decision-making [36], accuracy and learning (i.e., changes in accuracy over time) is assessed via correspondence to this ground truth signal.

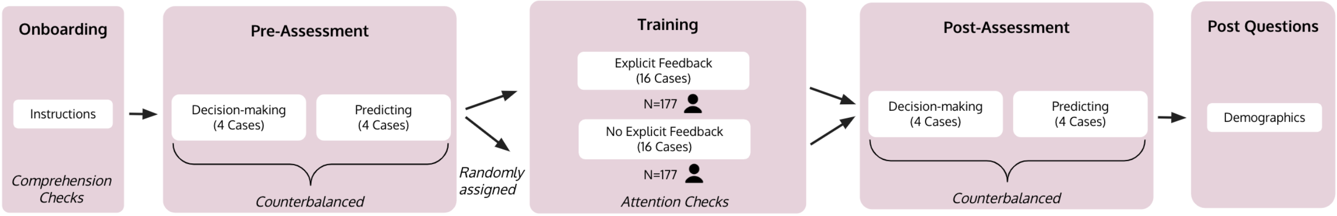


Figure 1: A high-level overview of the study design. The study included five phases: Onboarding, Pre-Assessment, Training, Post-Assessment, and Post Questions. In the Training phase, participants were randomly assigned between-subjects to the *Practice* or the *Practice + Explicit Feedback* case.

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

2.3 Research Questions

Domains involving human decision-making have long recognized that **realistic practice** and **tailored feedback** are key processes for learning. Literature from the learning sciences has demonstrated how practice improves performance across domains, from academic subjects like math and psychology to complex medical procedures and even to professionals' abilities to navigate challenging social interactions [22, 29, 34]. In prior literature, these simple practice effects have been substantially enhanced when supplemented with explicit feedback [17, 33, 34]. In real-world decision-making settings such as social work, healthcare, and education, where there is typically a long lag (on the order of months or even years) between decisions and corresponding outcomes, such simulation-based training approaches have the potential to greatly accelerate learning.

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

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

3 METHODS

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

3.1 Context

We ground our investigations in the context of child maltreatment screening: a complex, social decision-making context where the use of AI-based decision support tools is rapidly spreading [12, 50, 52]. The task design and data are drawn from a real-world ADS deployment: the Allegheny Family Screening Tool (AFST). As one of the longest deployed and most well-known ADS tools in social services, the AFST has influenced the design and deployment of many other AI-assisted decision-making tools across the U.S. [50]. The Allegheny County Department of Human Services (DHS) deployed the AFST in 2016 to assist hotline call screeners and supervisors in prioritizing alleged child maltreatment cases for investigation [12]. As is common for ADS tools deployed in complex, real-world settings, the AFST predicts readily measurable yet indirect *proxies* for the construct of interest to decision-makers. Whereas frontline decision-makers focus on ensuring children's immediate safety, the AFST predicts a proxy outcome on a much longer time horizon: a child's risk of being *placed into foster care* within the next *two years* [8, 30]. The AI-assisted child maltreatment context also involves information asymmetries: prior field studies find that frontline workers' decisions are informed by rich, qualitative information about a given case (e.g., reported allegations), which is unavailable to the ADS model [30, 53].

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

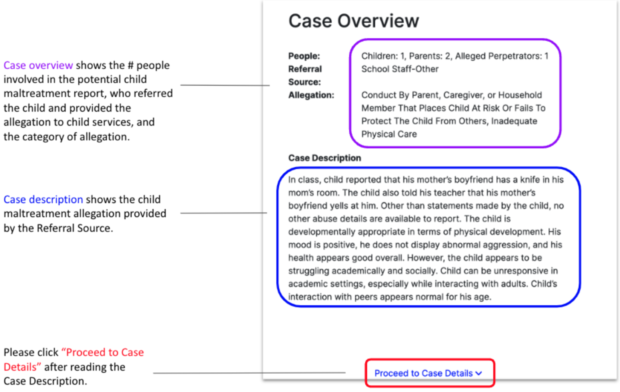
3.2 Study Design and Materials

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

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

The **pre- and post-assessments** were designed to assess both (a) how participants make child maltreatment screening decisions with AI assistance; and (b) participants' ability to predict what risk score the AI model will assign to a given case. Assessing participants' ability to predict AI outputs, both before and after the training phase enabled us to investigate whether repeated practice and feedback on AI-assisted decision-making would additionally improve their ability to mentally simulate the AI model's behavior [4, 40, 48]. Each assessment included a series of four AI Score Prediction activities and four AI-Assisted Decision-Making activities (which were identical to the decision-making activities presented during the training phase). During the assessment phases of the study, participants were not shown any explicit feedback on either their decisions or their AI score predictions, regardless of which experimental condition they were assigned to for the training phase. In each assessment, we counterbalanced the order in which the two types of assessment activities were shown. For each case shown during the assessment phases, participants were asked to elaborate (via open text) on how they made their AI-assisted decision or AI score prediction immediately after inputting their response.

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


Case overview shows the # people involved in the potential child maltreatment report, who referred the child and provided the allegation to child services, and the category of allegation.

Case description shows the child maltreatment allegation provided by the Referral Source.

Please click "Proceed to Case Details" after reading the Case Description.

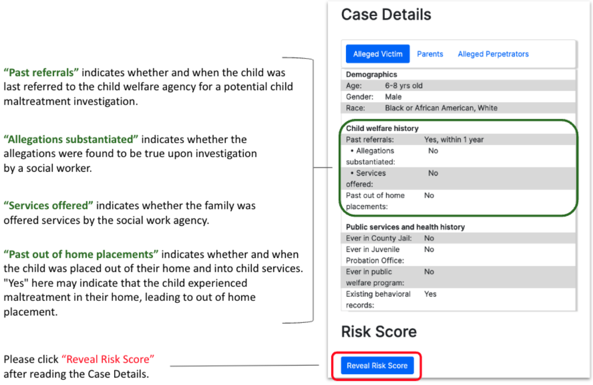
Case Overview

People: Children: 1, Parents: 2, Alleged Perpetrators: 1
Referral Source: School Staff-Other
Allegation: Conduct By Parent, Caregiver, or Household Member That Places Child At Risk Or Fails To Protect The Child From Others, Inadequate Physical Care

Case Description

In class, child reported that his mother's boyfriend has a knife in his mom's room. The child also told his teacher that his mother's boyfriend yells at him. Other than statements made by the child, no other abuse details are available to report. The child is developmentally appropriate in terms of physical development. His mood is positive, he does not display abnormal aggression, and his health appears good overall. However, the child appears to be struggling academically and socially. Child can be unresponsive in academic settings, especially while interacting with adults. Child's interaction with peers appears normal for his age.

Figure 2: A screenshot of the Case Overview information panel that provides an overview of the alleged child maltreatment case and a text box with the allegations as a qualitative description. The case content, including the allegations, are taken from real historical data recording past experienced workers' AI-assisted decision-making tasks.


"Past referrals" indicates whether and when the child was last referred to the child welfare agency for a potential child maltreatment investigation.

"Allegations substantiated" indicates whether the allegations were found to be true upon investigation by a social worker.

"Services offered" indicates whether the family was offered services by the social work agency.

"Past out of home placements" indicates whether and when the child was placed out of their home and into child services. "Yes" here may indicate that the child experienced maltreatment in their home, leading to out of home placement.

Please click "Reveal Risk Score" after reading the Case Details.

Case Details

Demographics

Age: 6-8 yrs old
Gender: Male
Race: Black or African American, White

Child welfare history

Past referrals: Yes, within 1 year
• Allegations substantiated: No
• Services offered: No
Past out of home placements: No

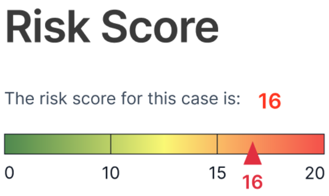
Public services and health history

Ever in County Jail: No
Ever in Juvenile Probation Office: No
Ever in public welfare program: No
Existing behavioral records: Yes

Risk Score

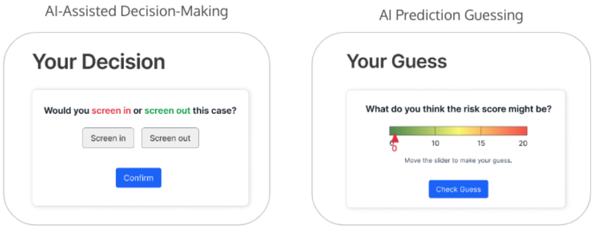
Figure 3: A screenshot of the Case Details information panel that provides public records and past referral information on the individuals in the alleged child maltreatment case. The content is taken from the same historical case information used to populate the Case Overview and AI Risk Score.

AI-Assisted Decision-Making Activity. To inform their decisions in the AI-assisted decision-making activity, participants were presented with both a referral vignette and an AI prediction. For each case, the AI prediction and content in the referral vignette was drawn and adapted from case notes on actual cases for which past workers subsequently made AI-assisted decisions, in the Allegheny County context. The referral vignette first shows a **Case Overview** which includes **qualitative allegations**. The child maltreatment allegations were typed by experienced workers based on calls they received (e.g., from a neighbor or a teacher). For each allegation drawn from the historical dataset, we removed any potentially identifying information (e.g., details pertaining to specific regions or households), while still preserving high-level properties of the allegations, as this information could be relevant to the interpretation of other case-related information, including the AI prediction. Figure 2 shows a screenshot of a **Case Overview** section for a sample referral vignette. Next, we showed **Case Details** (Figure 3) including demographic information, child welfare history, and public services and health history of the alleged child victim, parent(s), and alleged perpetrator(s). For the purposes of our experiment, we opted to show a smaller subset of demographic information than



The risk score for this case is: 16

Figure 4: A screenshot of the AI Risk Score interface shown to participants in the AI-assisted decision-making task. This mirrors the actual AFST interface used by social workers.



Would you screen in or screen out this case?

What do you think the risk score might be?

Move the slider to make your guess.

Figure 5: A screenshot of the question that participants are asked in the AI-assisted decision-making activity (left) and AI score prediction activity (right). To inform this decision, participants also see information from the Case Overview, Case Details, and (for the decision-making task only) the AI prediction.

actual workers are able to access in practice, both to exclude potentially identifying information and to ensure that participants would have sufficient time to look through the information (cf. [9]). Finally, we showed the **Risk Score** (Figure 4): the actual AFST score for the given historical case.

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

AI Score Prediction Activity. In the AI Score Prediction activity, participants were similarly shown referral vignettes (including the Case Overview and Case Details) that replicated the decision-making tasks of actual social workers using an ADS. However, rather than showing the AI Risk Score, participants were asked to make their own prediction on the AI output on a scale of 1 to 20. Figure 5 (right) includes a screenshot of the question participants were asked after seeing the referral vignette.

3.3 Feedback Design

Participants in the *Practice + Explicit Feedback* condition received immediate feedback on their decisions during the study's training phase. Given that we do not typically have access to a single, reliable "ground truth" label in the context of child maltreatment screening, as discussed above, we explore the learning effects of providing participants with multiple feedback signals in two primary forms: *observed outcomes* and *historical human decisions* (Figure 6). By showing multiple, imperfect "ground truths," we provide the human decision-maker with opportunities to reflect on disagreements amongst the ground truths and their own initial decisions, to learn how to more appropriately calibrate reliance on the ADS model over time.

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

(1) Whether the child was **placed within two years**: In the AFST context and in our study, this is the AI model's predictive target. If the child was eventually placed out of their home within two years, this may signal that the child was maltreated and should have been screened in and subsequently investigated. However, there are other unobservable factors unrelated to the screening decision that may have led to this outcome [13]. For example, the child's parents may suddenly pass away, requiring the child to be placed into foster care in the absence of any reliable grandparents or relatives. Thus, placement in two years is an imperfect proxy for the accuracy of the screening decision.

(2) Whether the child was **re-referred to the agency within two years**: Similar to (a), a range of other unobservable factors unrelated to the decision may have led to this outcome. For example, certain stakeholders (e.g., divorced parents doing "retaliation calls") may feel incentivized to call the hotline and alleged maltreatment, when in fact, the child is safe [30].

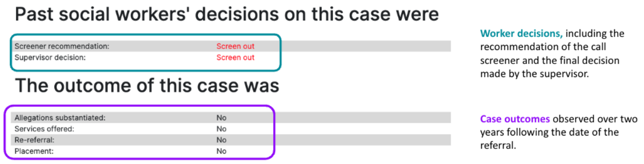
(3) Whether the **allegations were substantiated upon screen in**: Only cases that were historically screened in can observe this outcome. Allegations of potential child maltreatment being substantiated based upon investigation may indicate that the child was indeed maltreated, and thus screening in was the "right" decision. *But*, allegations, even if substantiated, may result from cultural misunderstandings or implicit biases from the caller [8].

(4) Whether **services were offered to the family upon screen in**: Only cases that were historically screened in can observe this outcome. A family that is offered services may indicate that the child was maltreated, and thus was offered services that can support recovery. In this case, a decision to screen in would be "right." *On the other hand,* even if a child was not maltreated, a family may receive services as a preventive measure to minimize risk of future maltreatment.

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

(1) Whether the **call screener's recommendation** was a screen in or screen out: This is the call screening recommendation a past experienced social worker made, based on the case information and AI risk score they saw.

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



Past social workers' decisions on this case were

Screener recommendation: Screen out
Supervisor decision: Screen out

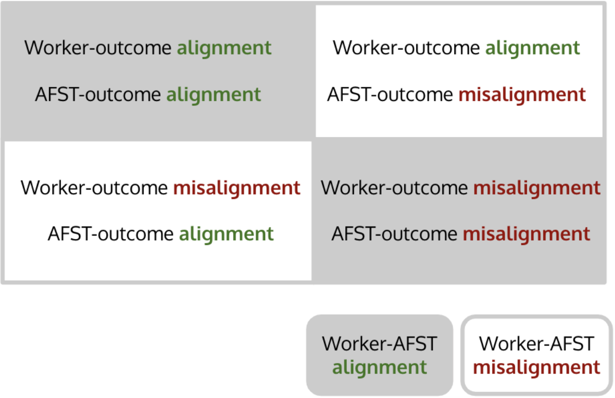
The outcome of this case was

Allegations substantiated: No
Services offered: No
Re-referral: No
Placement: No

Worker decisions, including the recommendation of the call screener and the final decision made by the supervisor.

Case outcomes observed over two years following the date of the referral.

Figure 6: A screenshot of the six imperfect proxies (screener recommendation, supervisor decision, allegation substantiation, services offered, re-referral, and placement) shown as feedback to participants in the AI-assisted decision-making task in the *Practice + Explicit Feedback* condition. Participants in the *Practice* condition were shown the next case immediately after submitting their decision.



Worker-outcome alignment AFST-outcome alignment	Worker-outcome alignment AFST-outcome misalignment
Worker-outcome misalignment AFST-outcome alignment	Worker-outcome misalignment AFST-outcome misalignment

Figure 7: A 2x2 confusion matrix breaking down the three alignment measures into four high-level categories.

Taken together, this set of feedback signals captures both the accuracy of the AI model’s predictions (with respect to the proxy outcome it is trained to predict), as well as what past experienced human workers believed was the right decision after reviewing both the case information and the AI prediction.

3.4 Case Selection

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

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

- **Worker-AFST alignment:** Cases where the worker agreed with the AFST decision[1] (i.e., both state screen in or screen out).
- **Worker-outcome alignment:** Cases where the worker’s assessment aligns with the observed outcome (i.e., screen in & placed or screen out & not placed).
- **AFST-outcome alignment:** Cases where the AFST’s assessment aligns with the observed outcome.

To guide the sampling of cases, we then constructed a 2x2 confusion matrix breaking down the three alignment measures into four categories as shown in Figure 7.

3.4.2 Case Sampling and Randomization. We randomly sampled cases from historical referral data according to our case categorization. We drew 6 cases from each of the 8 sub-categories, yielding a total of 48 cases. Cases were sampled with equal weighting across categories to enable us to study learning effects relative to historical worker and algorithm behavior. Cases within each category were then randomly assigned across the pre-assessment, training, and post-assessment phases of the study. Across participants, cases were assigned to the same study phase (i.e., pre-assessment, training, or post-assessment) to enable item-level analyses, but were presented in randomized order within each phase.

3.5 Participants and Recruitment

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

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

[1] We designate cases with an AFST score ≥ 15 as “screen in” and cases with an AFST score < 15 as “screen out”. This screen-in threshold aligns with the high risk cutoff shown to call screeners on the AFST tool interface [56]. Use of this threshold also follows prior analyses of automated AFST decisions [8]

implications of our recruitment criteria on the interpretation of our findings in the Discussion (Section 5.3).

The study included two comprehension checks and two attention checks that were used to exclude participants from the final dataset. The design of the comprehension and attention checks followed Prolific's policies [2]. Participants who failed to answer at least one of the comprehension checks were navigated to the end of the study. Participants who failed at least one of the attention checks were excluded from the final dataset.

3.6 Measurement and Analysis

Drawing upon findings from prior field studies [e.g., 8, 30], we defined a set of *indicators* of critical use in the context of AI-assisted child maltreatment screening, overviewed below.

Decision alignment measures. To quantitatively study changes in participants' decision-making, we defined two decision alignment measures, covering both the "human" and the "AI" portion of the past recorded AI-assisted decisions used in our study. Because our experiment uses real, historical data (as described in Section 3.4), this enables us to explore how participants' decisions compare to past decisions made by workers experienced in AFST-assisted decision-making. The first of our measures captures agreement with the *AI model*, while the other captures agreement with *experienced workers:*

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

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

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

We supplemented our pre-post analyses with analyses of participants' learning across the entire study. In particular, we fit mixed effects models to examine interactions between the practice opportunity number Z_{num} (ranging from 1 - 24) and the presence of explicit feedback $Z_{feedback}$ (a binary indicator). To analyze differences across case categories, we additionally examined interactions between the case type and the practice opportunity number. As described in Section 3.4.1, each case category applied to 50% of the 24 total cases, making it possible to run these analyses on balanced amounts of data.

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

3.6.1 Standard accuracy measure: Proxy-based accuracy.

The standard measure that is used in existing research literature to measure appropriate reliance is accuracy with respect to the *model's predictive target* [21], which we call Proxy-based Accuracy. We say there are *increases* in **Proxy-based Accuracy** if participants' decisions become more accurate with respect to the model's predictive target. For a given case, a decision is *accurate* with respect to the model's predictive target if the participant either screened in a case that resulted in out-of-home placement in two years or screened out a case that did not result in out-of-home placement in two years.

3.7 Positionality

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

Importantly, we explore training towards *critical use* in this paper because we believe doing so is crucial in many settings where AI systems are deployed. As we have argued above, AI performance in complex, real-world settings is bound to be imperfect, and some limitations are to be expected—including some degree of target-construct mismatch and limitations in the information a model is able to access or interpret. Training towards critical use may mitigate downstream negative impacts of AI model limitations. However, we advise such training only in settings where the deployment of an AI-based decision support tool is actually justifiable in the first place. Training alone *cannot* overcome fundamental design flaws in the AI model or surrounding social systems. While we examine AI-assisted child maltreatment screening as an example of a highly complex social decision-making setting in this study, we note that existing AI deployments in this domain have seen many fundamental design critiques, including prior research by members of our team [8, 13, 18, 30, 31, 53, 57].

4 RESULTS

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

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

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

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 1: Coefficient estimates for models examining pre-post learning effects. Each cell shows the coefficient estimate followed by the standard error for a given term (rows) and model (columns). A positive *Phase* coefficient indicates an increase in the measure in the post-assessment as compared to the pre-assessment baseline. A visual summary of these findings is reported in Table 2.

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

Strikingly, we found that participants learned to disagree with AI predictions more often, in ways that aligned more with past frontline workers' decisions, even in the *Practice* condition, where participants did not receive any explicit feedback about past frontline workers' decisions. An examination of participants' open-text feedback suggests that this occurred because participants learned through practice to integrate case-specific, qualitative details from the allegations when making their decisions, with a focus on assessing nearer-term risks to the child's safety (in contrast to the AI model's focus on predicting a proxy outcome, placement, on a longer time window). Past research suggests that this mirrors the way experienced workers make decisions in the field [8, 30]. In turn, participants' decisions began to diverge from model predictions and resembled those of past experienced workers, with increased practice (Section 4.4).

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

Moreover, receiving explicit feedback on decisions did not impact participants' learning with respect to *decision-making*, compared to the effects of repeated practice alone. Our analyses suggest that the qualitative case narratives present in both conditions are a rich information source through which participants can learn to calibrate their reliance on AI predictions. Thus, even in the absence of *explicit* feedback on decisions: a training interface that provides accelerated practice, with opportunities to cross-check AI predictions against qualitative narratives, appears to be a stronger baseline than expected.

In the following subsections, we first present findings from quantitative analyses of participants' learning and decision-making (Sections 4.1, 4.2, 4.3, 4.5). In Section 4.4, we present complementary findings from a qualitative examination of participants' decision-making *processes* across practice opportunities. Table 1 includes aggregate outcome-based findings, and Table 2 provides a summary of the directionality of outcome-based findings across the two conditions.

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

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

Participants started the study by relying on the AI prediction more often, including in cases where the past experienced worker disagreed with the AI prediction. But over time, participants *were less likely to get nudged by the AI prediction.* Instead, over time, participants began to disagree with the AI prediction more often

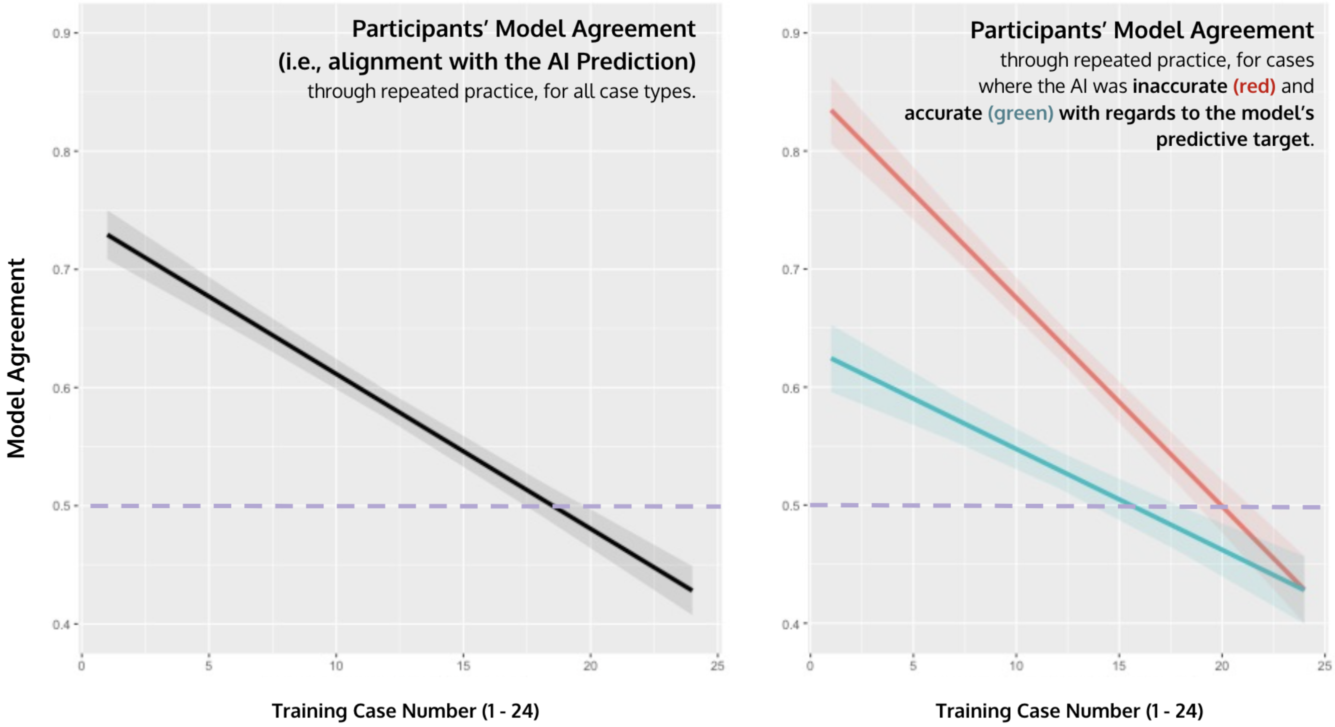


Figure 8: Mixed effects regression visualizing participants' agreement with the AI prediction, across the 24 training cases. Shaded regions indicate standard error. The purple dotted line indicates average accuracy of the AI model with respect to its (proxy) predictive target.

for all case types. As shown in Figure 8, in the first case, over 70% of the participants agreed with the AI prediction. However, in the 24th case, only 42% of the participants agreed with the AI prediction. This finding was corroborated by the learning effects regression (Table 1), which showed a significant decrease in Model Agreement between pre- and post-assessments as indicated by the study phase coefficient (Coef.=-0.407; p<.001; 95% CI: [-.46, -.35]).

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

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

With practice, participants' decisions began to align more with the final decisions of past experienced workers. Overall, Figure 9 (middle) shows that, at the start of the training, 52% of the participants made decisions that were accurate with respect to the past experienced worker. By the end of the training, 61% of the participants made decisions that were accurate with respect to the past experienced worker. This finding is consistent with learning effects regression results, which show a significant increase in Worker-based Accuracy scores between pre- and post-assessments (Coef.=0.083; p<.01; 95% CI: [0.027, 0.140]).

With increased practice, participants' decisions were especially likely to be accurate with respect to the past experienced worker when *disagreeing* with the AI prediction. As shown in Figure 9, for the category of cases in which the past experienced worker disagreed with the AI prediction (right), for the first case, only 30% of participants made decisions that were accurate with respect to the past experienced worker and in disagreement with the AI prediction. By the 24th case, nearly 70% of participants made decisions that were accurate with respect to the past experienced worker and in disagreement with the AI prediction. On the other hand, for the category of cases in which the past experienced worker agreed with the AI prediction, at the start of the training, nearly 75% of participants made decisions that were accurate with respect to the past experienced worker. By the 24th case, this proportion went down to just over 50% of participants.

Additionally, participants' decisions became more accurate with respect to the past experienced worker regardless of whether those past experienced workers made decisions that were inaccurate or accurate with respect to the model's predictive target, an accuracy-based measure that is widely used in prior studies evaluating the accuracy of AI-assisted decision-making with this AI model, the AFST (e.g., in [8, 14]). As shown in Figure 9, for cases in which the past experienced worker was inaccurate with respect to the

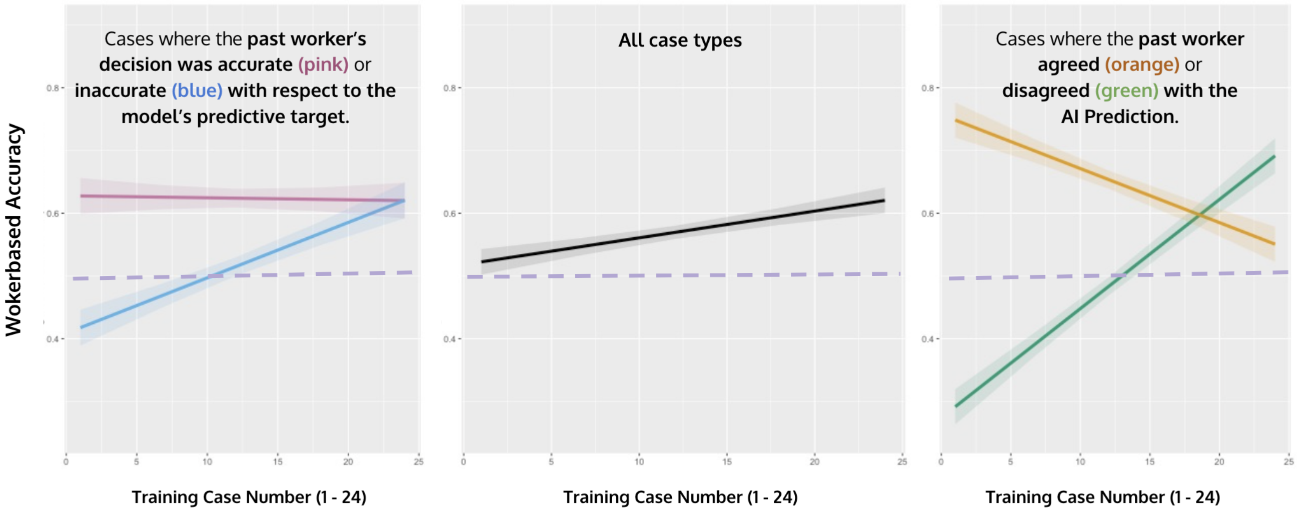


Figure 9: Mixed effects regression visualizing participants' learning across repeated practice, with respect to the worker-based accuracy measures (i.e., accuracy with respect to past experienced workers).

model's predictive target (left), around 60% of participants at both the start and end of the training made decisions that were accurate with respect to the past experienced worker. However, for cases in which the past experienced worker was accurate with respect to the model's predictive target, only around 40% of participants made decisions that were accurate with respect to the past experienced worker in the first case. Yet, by the 24th case, this number increased to around 60% of cases, erasing any differences in decision behavior for cases in which past experienced workers were accurate versus inaccurate with respect to the model's predictive target. As shown in Table 1, explicit feedback did not have a significant effect on Worker-based Accuracy.

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

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

While disagreeing with the AI prediction more often with time, participants were simultaneously *getting better at guessing what the AI prediction would be on a given case.* Participants who received practice alone, compared to those who additionally received feedback on their decisions, saw somewhat more improvement in their guessing performance. As shown in Table 1, there was a significant decrease in guess mean squared error in the post-assessment as compared to the pre-assessment (Coef.=-3.03; p<.001; 95% CI: [-3.45, -2.60]). Additionally, participants assigned to the *Practice* condition saw a greater improvement in their guessing performance as compared to participants in the *Practice + Explicit Feedback* condition (Coef.=0.608; p<.05; 95% CI: [0.05, 1.15]). Overall, this result indicates that, despite disagreeing with predictions, participants in both conditions were actively engaging with the learning activity and improving their skills at mentally simulating the behavior of the model. However, participants who did *not* receive explicit feedback on their AI-assisted decisions improved these skills more. As we discuss in Section 5.2, it is possible that, compared to the explicit feedback, the information reported in the case (including the qualitative case narratives) served as a more powerful signal through which to learn how the AI model behaves on different cases, compared to the explicit feedback signals provided. It is possible that the *Practice* condition saw greater improvement because explicit feedback on AI decisions distracted participants from learning to predict model behavior based on qualitative narratives.

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

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

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

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

Table 2: Visual summary of model findings reported in Table 1. Upwards arrow (rose-colored cell) indicates increases in the measure value after the training. Downwards arrow (blue cell) indicates decreases in the measure value after the training, and two downwards arrows (darker blue cell) indicates greater decreases in the measure value in the repeated practice condition compared to when explicit feedback was shown alongside repeated practice.

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

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

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

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

4.4.2 Making decisions with the goal of ensuring immediate safety. Beyond leveraging their unique ability to access and integrate context-specific information when making decisions, the goals that participants had when making decisions may also explain the learning effects. In particular, it is possible that participants learned to disagree with the AI prediction, while learning to make decisions like past experienced workers, partly because their own decision-making goals aligned more strongly with those of the past experienced workers. At the end of the experiment, participants

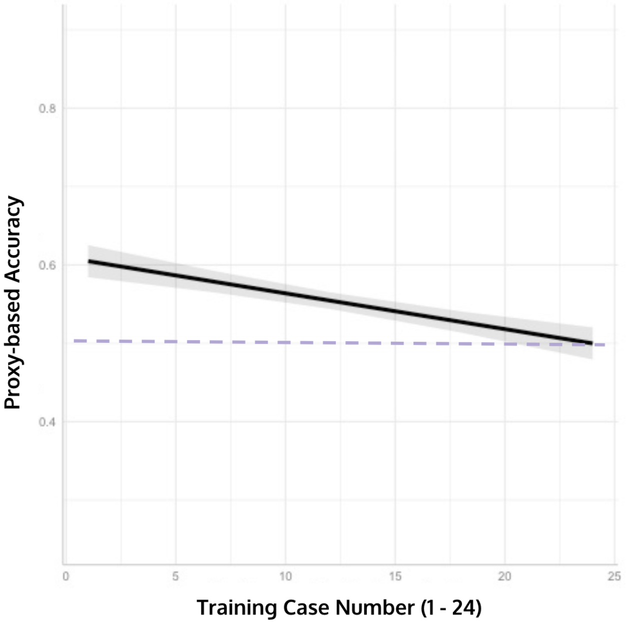


Figure 10: Participants’ Proxy-based Accuracy measures against the baseline accuracy of the model (in purple).

were asked to describe their own goals when making child maltreatment screening decisions. Many participants described the importance of considering near-term safety risks or harms to the child, which is misaligned with the longer-term outcomes that the AI model is trained to predict. Both participants with and without domain knowledge in social work described goals related to “immediate safety and harm,” aligning with the goals documented of the actual past experienced social workers in this domain. For example, one social worker described: “I am looking for immediate safety concerns first. Then making sure physical needs are being met. If either of these are a concern, it is an immediate screen in.” Another social work graduate student described the importance of “current” evidence: “I assess first for evidence of maltreatment or harm currently going on in the home and then consider if additional investigation is necessary to rule out possible abuse or neglect, I then consider if there is enough "probable cause" to truly investigate this further.” A participant without social work domain expertise similarly described: “My goals were whether or not there would be any immediate danger posed to the child either through their environment or by their parents.”

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

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

Participants gradually began to disagree with the AI prediction regardless of whether the AI was inaccurate or accurate with respect to the model’s predictive target. As shown in Figure 8, for the first case in which the AI was accurate with respect to the model’s predictive target, 80% of participants agreed with the AI prediction. For the first case in which the AI was inaccurate with respect to the model’s predictive target, 60% of participants agreed with the AI prediction. However, in the 24th case, only 40% of participants agreed with the AI prediction, regardless of whether the AI prediction was accurate or inaccurate with respect to the model’s predictive target.

Importantly, this gradual disagreement with the AI prediction occurred *regardless of whether the participant received feedback on their decisions.* As shown in Table 1, there is no feedback effect on participants’ decisions to agree or disagree with the AI prediction over time (Coef. 0.02, p < .50, 95% CI: [-0.03, 0.09]).

5 DISCUSSION

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

5.1 Evidence of *critical use* from training

Our findings suggest that training improved *critical use* of the AI tool—humans’ ability to situate AI predictions against potentially complementary knowledge available to them (but not the AI model). A body of prior research on AI-assisted child maltreatment screening has documented how social workers draw on their knowledge of qualitative, contextual information to calibrate their reliance upon AI-based decision support (e.g., [30, 51, 53]). In doing so, experienced workers can overcome some of the challenges posed by target-construct mismatch and limitations in the information available to AI models, mitigating erroneous AI outcomes and reducing disparities in decisions [8, 14]. In our study, we found that **participants learned to calibrate their reliance upon AI predictions in ways that resembled experienced workers.** In particular, with increased practice, participants learned to disagree more often with AI predictions (Section 4.1) and, in turn, came to make decisions that aligned more with those of past workers with extensive experience making AI-assisted decisions (Section 4.2). Moreover, through practice making AI-assisted decisions participants improved their ability to mentally simulate the AI model and predict how it would behave on specific cases (Section 4.3). Our qualitative examination of participants’ explanations suggest that participants drew upon knowledge regarding the actual decision-making goals (i.e., screening for shorter-term safety) and interpreted qualitative narratives unavailable to the AI model in order to inform their AI-assisted decisions (Section 4.4) and learn how the AI model behaves on different kinds of cases. However, we note that future work is needed to understand the *extent* to which specific downstream impacts of target-construct mismatch and information asymmetries (e.g., systematic errors and unfairness in decision-making) can be mitigated through training.

Analyses using a standard measure of accuracy for human-AI decision-making would indicate that increased practice led participants to make *less* accurate decisions, when accuracy is measured with respect to the model's target proxy (Section 4.5). Interpreting proxy-based accuracy in isolation from the rest of our analyses may lead to the conclusion that participants are simply falling into patterns of algorithm aversion [5, 15]. However, in analyzing indicators of *critical use*, including both process- and outcome-oriented signals, our findings suggest that participants disagreed with AI outputs in sophisticated ways. Indeed, in complex, real-world decision-making settings, common outcome-focused metrics for evaluating AI-assisted decisions—for example, comparing human decisions with the AI model's ground truth label—can provide an incomplete understanding of the quality of each decision [21, 30]. Our findings point to a compelling opportunity to improve evaluations of AI-assisted decision-making in other complex domains, by designing and employing context-specific measures of *critical use, in addition to examining proxy-based accuracy.* Moreover, to expand the set of indicators that could be used to measure decision-making, future work should innovate on ways to better capture process-based signals for decision quality (e.g., [35, 42]). We further discuss how our study findings and approach may generalize across domains in Subsection 5.4.

Limitations. For the purposes of our study, we selected cases where the AI tool had an average proxy-based accuracy of (50%). This is lower than the the AFST model's actual accuracy with respect to its proxy. Our decision to include cases that aggregated to 50% proxy-based accuracy of the AI tool was informed by our goal of having a well-balanced set of examples to conduct case type-specific analyses and of prior literature discussing the risks of over-reliance in the absence of sufficient examples of the AI model erring [47]. While we cannot rule out that this contributed to increased disagreement with the AI tool, our findings indicate that it is *not sufficient* to explain participants' learning. For example, participants aligned less with the model proxy following training even in the *Practice* condition, when participants did not receive explicit feedback on the accuracy of the AFST with respect to its targeted proxy (see [41]).

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

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

Prior work from the learning sciences suggests that for certain learning tasks, providing learners with "grounded" feedback—concrete representations that offer rich, meaningful signals about whether or not an individuals' targeted outcome is achieved—is more effective than simply telling learners whether a response is correct or incorrect. For example, when teaching students to perform algebraic transformations, showing students a graphical representation of equations they enter can help students immediately *recognize for themselves* whether an equation is likely to be correct or incorrect[58]. Future work on training for AI-assisted decision-making should further explore the design space of rich, grounded feedback mechanisms that can help humans (learn to) calibrate their reliance. In particular, whereas our study provided explicit feedback in the form of categorical outcomes and workers' final decisions, future work should explore what forms of grounded feedback could support more effective process-oriented learning in a given context. In complex decision-making settings such as child maltreatment screening, observable outcomes and even final decisions) are noisy signals for decision making quality [8, 31]. Therefore, rather than providing human decision-makers with outcome-based feedback that attempt to directly "tell" a learner whether a given decision was accurate, providing rich, grounded feedback may make it easier to engage learners' sensemaking and help them *assess for themselves* whether and why a decision may be right or wrong. Training that focuses on better scaffolding the decision-making process—for example, by pairing a novice decision-maker with an expert decision-maker to collaboratively reason about each decision—may improve critical thinking and sensemaking processes that also allow humans to better calibrate reliance on ADS outputs.

5.3 On the absence of domain expertise effects

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

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

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

5.4 Generalizability of design decisions and findings

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

ACKNOWLEDGMENTS

We acknowledge support from Toyota Research Institute (TRI), the UL Research Institutes through the Center for Advancing Safety of Machine Intelligence (CASMI) at Northwestern University, the Carnegie Mellon University Block Center for Technology and Society under Award No. 53680.1.5007718 and 55410.1.5007719, the National Science Foundation (NSF) under Award No. 1939606, 2001851, 2000782, and 1952085, and two awards from the National Science Foundation Graduate Research Fellowship Program (GRFP).

REFERENCES

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

[2] 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*. 1–16.

[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] Jason W Burton, Mari-Klara Stein, and Tina Blegind Jensen. 2020. A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making* 33, 2 (2020), 220–239.

[6] 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.

[7] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2021. Onboarding Materials as Cross-functional Boundary Objects for Developing AI Assistants. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.

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

[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. *Conference on Human Factors in Computing Systems - Proceedings* (2021), 1–17. https://doi.org/10.1145/3411764.3445308

[10] Lingwei Cheng and Alexandra Chouldechova. 2022. Heterogeneity in Algorithm-Assisted Decision-Making: A Case Study in Child Abuse Hotline Screening. *arXiv preprint arXiv:2204.05478* (2022).

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

[12] 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

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

[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. *arXiv* (2020), 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] Kate Donahue, Alexandra Chouldechova, and Krishnaram Kenthapadi. 2022. Human-Algorithm Collaboration: Achieving Complementarity and Avoiding Unfairness. *arXiv preprint arXiv:2202.08821* (2022).

[17] Gilan M El Saadawi, Roger Azevedo, Melissa Castine, Velma Payne, Olga Medvedeva, Eugene Tseytlin, Elizabeth Legowski, Drazen Jukic, and Rebecca S Crowley. 2010. Factors affecting feeling-of-knowing in a medical intelligent tutoring system: the role of immediate feedback as a metacognitive scaffold. *Advances in Health Sciences Education* 15, 1 (2010), 9–30.

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

[19] Krzysztof Z Gajos and Lena Mamykina. 2022. Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. In *27th International Conference on Intelligent User Interfaces*. 794–806.

[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] Luke Guerdan, Amanda Coston, Zhiwei Steven Wu, and Kenneth Holstein. 2023. Ground (less) Truth: A Causal Framework for Proxy Labels in Human-Algorithm Decision-Making. *arXiv preprint arXiv:2302.06503* (2023).

[22] Anne Helsdingen, Tamara Van Gog, and Jeroen Van Merriënboer. 2011. The effects of practice schedule and critical thinking prompts on learning and transfer of a complex judgment task. *Journal of Educational Psychology* 103, 2 (2011), 383.

[23] Patrick Hemmer, Max Schemmer, Niklas Kühl, Michael Vössing, and Gerhard Satzger. 2022. On the Effect of Information Asymmetry in Human-AI Teams. *arXiv preprint arXiv:2205.01467* (2022).

[24] Kenneth Holstein and Vincent Aleven. 2022. Designing for human–AI complementarity in K-12 education. *AI Magazine* 43, 2 (2022), 239–248.

[25] 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.

[26] Kenneth Holstein, Maria De-Arteaga, Lakshmi Tumati, and Yanghuidi Cheng. 2023. Toward supporting perceptual complementarity in human-AI collaboration via reflection on unobservables. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–20.

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

[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] Sarah J Kaka, Joshua Littenberg-Tobias, Taylor Kessner, Anthony Tuf Francis, Katrina Kennett, G Marvez, and Justin Reich. 2021. Digital simulations as approximations of practice: Preparing preservice teachers to facilitate whole-class discussions of controversial issues. *Journal of Technology and Teacher Education* 29, 1 (2021), 67–90.

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

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

[32] 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.

[33] Kenneth R Koedinger and Vincent Aleven. 2007. Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review* 19, 3 (2007), 239–264.

[34] Kenneth R Koedinger, Albert T Corbett, and Charles Perfetti. 2012. The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science* 36, 5 (2012), 757–798.

[35] Kwadwo Kyeremanteng and Gianni D'Egidio. 2015. Why process quality measures may be more valuable than outcome measures in critical care patients. *Biology and Medicine* 7, 2 (2015), 1.

[36] 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).

[37] Vivian Lai, Han Liu, and Chenhao Tan. 2020. " Why is 'Chicago' deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

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

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

[40] Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[41] Zhuoran Lu and Ming Yin. 2021. Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[42] William E McAuliffe. 1979. Measuring the quality of medical care: process versus outcome. *The Milbank Memorial Fund Quarterly. Health and Society* (1979), 118–152.

[43] Hussein Mozannar, Arvind Satyanarayan, and David Sontag. 2021. Teaching Humans When To Defer to a Classifier via Examplars. *arXiv preprint arXiv:2111.11297* (2021).

[44] Sendhil Mullainathan and Ziad Obermeyer. 2019. *A machine learning approach to low-value health care: wasted tests, missed heart attacks and mis-predictions.* National Bureau of Economic Research.

[45] Sendhil Mullainathan and Ziad Obermeyer. 2021. On the inequity of predicting a while hoping for B. In *AEA Papers and Proceedings*, Vol. 111. 37–42.

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

[47] Samir Passi and Mihaela Vorvoreanu. 2022. Overreliance on AI: Literature review. (2022).

[48] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. (2021). arXiv:1802.07810

[49] Charvi Rastogi, Liu Leqi, Kenneth Holstein, and Hoda Heidari. 2022. A unifying framework for combining complementary strengths of humans and ML toward better predictive decision-making. *arXiv preprint arXiv:2204.10806* (2022).

[50] 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

[51] Devansh Saxena, Karla Badillo-Urquiola, Pamela 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. *arXiv* 5, October (2021). arXiv:arXiv:2107.03487v2

[52] 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.

[53] Devansh Saxena, Seh Young Moon, Dahlia Shehata, and Shion Guha. 2022. Unpacking Invisible Work Practices, Constraints, and Latent Power Relationships in Child Welfare through Casenote Analysis. In *CHI Conference on Human Factors in Computing Systems*. 1–22.

[54] Treasury Board Secretariat. 2020. Directive on automated decision-making. *Ottawa (ON): Government of Canada (modified 2019-02-05* (2020).

[55] Eliane Stampfer, Yanjin Long, Vincent Aleven, and Kenneth R Koedinger. 2011. Eliciting intelligent novice behaviors with grounded feedback in a fraction addition tutor. In *Artificial Intelligence in Education: 15th International Conference, AIED 2011, Auckland, New Zealand, June 28–July 2011 15*. Springer, 560–562.

[56] 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).

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

[58] Eliane S Wiese and Kenneth R Koedinger. 2017. Designing grounded feedback: Criteria for using linked representations to support learning of abstract symbols. *International Journal of Artificial Intelligence in Education* 27 (2017), 448–474.

[59] Bryan Wilder, Eric Horvitz, and Ece Kamar. 2020. Learning to complement humans. *arXiv preprint arXiv:2005.00582* (2020).

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

[61] Qiaoning Zhang, Matthew L Lee, and Scott Carter. 2022. You Complete Me: Human-AI Teams and Complementary Expertise. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–28.