Submitted to *Management Science*
manuscript (Please, provide the manuscript number!)

Human-Artificial Intelligence Teaming and Effects of System Workload on the Screening of Child Maltreatment Reports

Child welfare organizations regularly receive a significant number of calls alleging child neglect or abuse. Due to limited resources available for investigations and services, it is crucial to accurately assess and screen these allegations before further investigation or intervention to ensure service quality and efficiency. Furthermore, investigations initiated based on unsubstantiated allegations can lead to harmful consequences for the family involved. To aid these essential screening decisions and enhance overall efficiency, a Predictive Risk Model (PRM), essentially an artificial intelligence (AI) tool, has been deployed by our research partner. We empirically investigate this human-AI collaboration in the call screening decision making, particularly examining the influence of workload on this collaboration. We find that human agents are more likely to deviate from the AI recommendations when the workload is either high or low. While the AI tool does not adjust for varying workloads, human agents seem to factor in the workload when making screening decisions. Specifically, they lean toward admitting more low-risk cases when the system load is low and lean toward rejecting more high-risk cases when the load is high, resulting in a U-shape relationship between deviation and workload. These findings indicate that human workers are able to complement the AI's recommendations by taking into account vital operational factors, such as system load. We provide evidence supporting that a sustainable investigation load is likely desirable for downstream operations, as increased investigation load is correlated with delayed investigation completion.

Key words: Child welfare operations, human-AI teaming, workload management, service system design.

1. Introduction

1.1. Child welfare services

The Centers for Disease Control and Prevention estimates that at least 1 in 7 children in the US have experienced child abuse or neglect in the last few years. In 2020, 1,750 children in the US died of abuse and neglect; and the total lifetime economic burden of child abuse and neglect was estimated to be $592 billion in 2018. This economic burden rivals the cost of other high-profile

public health problems, such as heart disease and diabetes. Abused or neglected children may suffer immediate physical injuries and are more likely to suffer emotional and psychological problems later in their lives, such as anxiety or post-traumatic stress. In the long term, maltreated children are at higher risk of future violence victimization and perpetration, substance abuse, delayed brain development, lower educational attainment, and limited employment opportunities CDC (2023).

Child welfare organizations are tasked with protecting children from abuse and neglect; their responsibilities include investigating child maltreatment allegations and providing services to children and families in need. However, not all allegations are substantiated, and unnecessary investigations may harm the families involved. Moreover, many child welfare organizations need to prioritize scarce resources and efforts to substantiated and serious referrals. Therefore, it is crucial that allegations are carefully screened before initiating formal investigations for the overall welfare of children and families.

A child maltreatment report or call is designated as a *referral*; one referral may involve more than one child and several allegations. Referrals are categorized into two types: child protective service (CPS) and general protective service (GPS). CPS reports are made by *mandated reporters*, i.e., adults who are required by law to report suspected child neglect or abuse. Mandated reporters are those adults who work or volunteer with children, including school employees, healthcare professionals, foster parents, and employees at other public services organizations Pennsylvania DHS (2024). Many US states (e.g., Michigan, Pennsylvania) established laws that mandate CPS referrals to be investigated within 24 hours Michigan HHS (2023), Pennsylvania DHS (2023a). In contrast, GPS referrals are only investigated by an assigned caseworker if *screened in*—accepted for investigation. Child abuse may take various forms; common abuse types are physical abuse, sexual abuse, emotional abuse (i.e., behaviors that harm a child's self-worth or emotional well-being, e.g., name-calling, shaming, rejecting, withholding love, and threatening), and neglect physically or emotionally CDC (2023).

1.2. Operations in a child welfare organization

We partner with a child welfare organization (CWO) in a northeastern US county. Figure 1 illustrates the workflow of the CWO. Specifically, their operations have three main stages: screening, investigation, and intervention. In the screening stage, all incoming GPS calls or reports are screened by call screeners in the CWO's *intake office*. Social workers at the intake office are mainly tasked with answering calls, assessing referrals, and investigating CPS referrals. GPS referrals, if screened-in, will be investigated by one of the five regional offices located in five different geographic regions in the county. CPS referrals are mandatorily investigated by social workers, often at the intake office. When an investigation concludes a referral is in need of services, a *case* for the referral

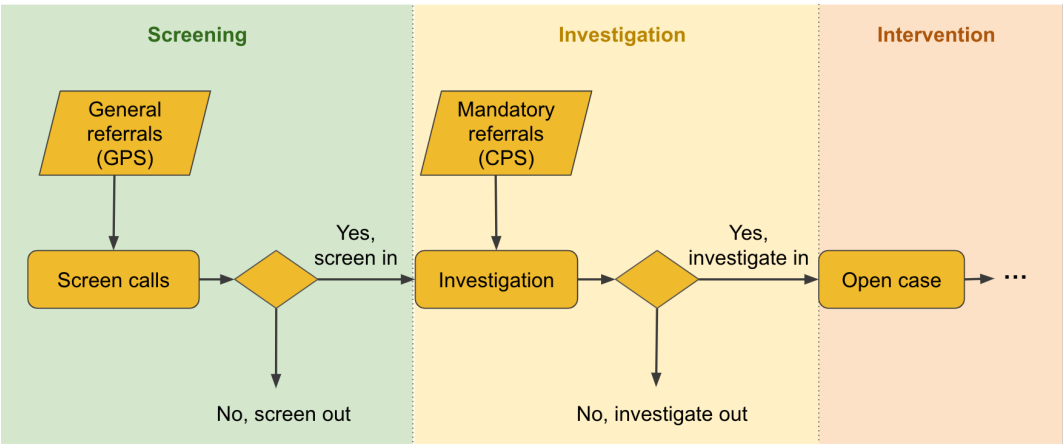

Figure 1 CWO workflow diagram.

is opened, and a social worker is assigned to intervene, offer protection, provide service, and work up a long-term solution for the children and family involved.

A predictive risk model (PRM) has been designed and implemented to enhance the screening decision making process within the CWO's child welfare system. The PRM harnesses the power of hundreds of data elements to generate the *PRM score*, which quantifies the likelihood of a child being placed *out-of-home* (OOH): The PRM score takes integer values ranging from 1 to 20, with a higher value indicating a higher probability of OOH placement within two years Pennsylvania DHS (2023b). The PRM is designed to complement clinical judgment by offering additional vital information to aid child welfare workers in making informed call-screening decisions. The PRM AI tool is generated and viewed only by call screeners; caseworkers who conduct investigations and provide services are not able to see any information from PRM.

Stage 1: Screening. Figure 2 illustrates the workflow of the screening stage. If someone in the community has concerns about suspected child maltreatment, they may report it to the CWO's hotline. When a call comes in, a call screener at the intake office answers the call while recording information into the computerized system following the established protocols. For all calls—those falling within a protocol or not—the screener runs the PRM on their computer, which will take the input of historical data (including demographics and past interactions with the CWOs) and the call information, and output the assessed risk score. After obtaining the PRM risk score, a field screen might be conducted if the call screener would like to gather more information about the children, the household, and the allegations for screening decision [1].

[1] A field screen is typically conducted if one or more of the following conditions are met: a) The child maltreatment report involves children three years old and younger who are directly impacted by the allegations. (b) If a report is the fourth referral associated with the same household within two years, yet there has not been any previous investigation into the household. And (c) a report involving children who receive education (through homeschooling, distance learning, or remote learning) at home.

Some referrals with specific characteristics (i.e., having a PRM score greater than 17 and involving a child aged 16 or under) fall into the *high-risk protocol* and, therefore, are designated to be screened in. In contrast, a small percentage of referrals fall into the *low-risk protocol* (i.e., having a PRM score no greater than 11 and no children under 12) and are encouraged to be screened out. Many other referrals fall into neither high-risk nor low-risk protocols. While the risk protocols guide screening decision-making, intake office supervisors can override the protocol-suggested screening outcomes at their discretion, given that they complete the override documentation. However, the computer information system shows the PRM score to the call screener only when a referral does not follow either risk protocol. For a referral that follows a high-risk protocol, the human-system interface displays "High-Risk Protocol: High Risk and Children Under 16 on Referral." Similarly, for a referral that follows a low-risk protocol, the human-system interface displays "Low-Risk Protocol: Low Risk and All Children Aged 12+ on Referral" and "recommended screen out."

Call screeners recommend screening in or screening out referrals. based on their assessment, the risk protocols or PRM risk scores, and field screens when necessary All call screeners' recommendations go through their supervisors, who make the final screening decisions based on the risk protocols, PRM scores, relevant data, and call screener recommendations. The call screeners and supervisors often discuss the evidence and screening rationale and then jointly make the screening decisions. Currently, 48.05% of incoming referrals are screened in for investigation.

Stage 2: Investigation. All screened-in GPS referrals and occasionally some CPS referrals are investigated by one of the five regional offices in different geographical regions within a US county. The intake office investigates most CPS referrals. Caseworkers must conclude any investigation within 60 days, though best practice encourages a 30-day completion of any CPS investigation. The investigation determines whether a CPS report is founded (i.e., there is a court action), indicated (i.e., there is substantial evidence that maltreatment occurred), unfounded (i.e., existing evidence does not meet the criteria for maltreatment), or pending (i.e., the CWO investigation cannot be concluded in 60 days because criminal or juvenile court action is initiated). GPS investigations conclude with whether allegations are valid or not based on collected evidence. If the caseworker(s) and their supervisor(s) believe there is an ongoing risk of child maltreatment in the household, then the referral may be accepted for service, and a child welfare case will be opened. If a child welfare case is not opened for the family, other community-based resources might still be offered to assist the household if needed.

Stage 3: Intervention. Upon opening a case (or equivalently, being accepted for service), a child welfare case worker will arrange a conference meeting with the family and their identified support (e.g., friends and other family members). At the conference meeting, they will discuss the family goals and devise a plan for services so the child can remain safely within the household.

Subsequent meetings are held with the same stakeholders to ensure that the family is making acceptable progress toward the goals. The investigation caseworker who conducts the investigation often remains with the family to provide service.

The CWO and other supervising bodies periodically review ongoing cases. In scenarios where they believe a child can no longer safely remain in a household, an out-of-home placement or removal from the home will be considered. Other interventions and services the CWO provides include foster care placement, adoption and permanent legal custodianship, and reunification.

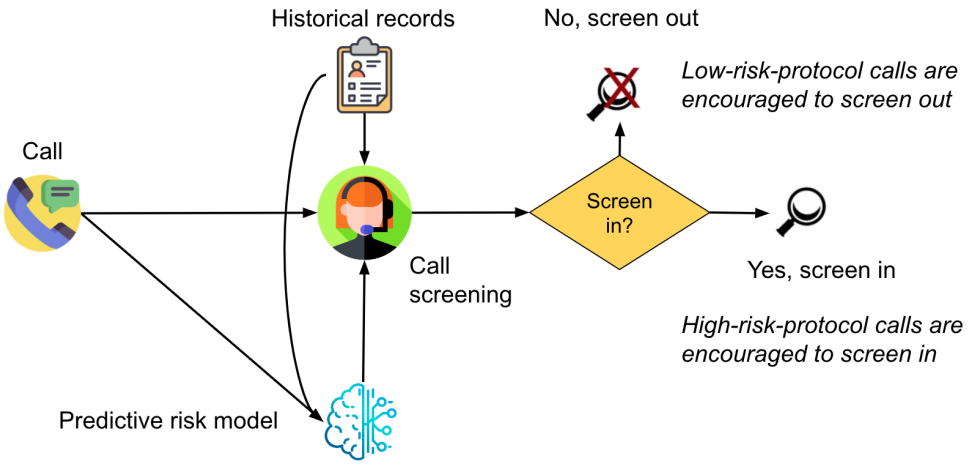

Figure 2 Workflow diagram: the screening step.

1.3. Research overview

In this project, we examine the interplay between human workers and the PRM, specifically assessing the influence of system load on child welfare screening recommendations and decisions. Recent research underscores the benefits of human-AI collaboration in diminishing disparities and errors. For example, Fogliato et al. (2022) highlights how human discretion in high-stakes contexts like child welfare can counterbalance algorithmic inaccuracies and reduce disparities. However, operational challenges have received scant attention in the existing literature. Our study fills the gap: We explicitly focus on the role of system load in human-AI collaboration in a high-intensity service environment with capacity constraints. Empirical evidence suggests that human workers are more likely to deviate from load-agnostic PRM recommendations when the workload level is either very high or low. Our findings reveal that human workers adeptly recognize the implications of system load on both organizational efficiency and service quality, thereby making screening choices, informed by clinical judgement and PRM's risk scores, to ensure workload sustainability.

Section 4 presents our main analysis results. Section 5 explores the impact of workload in the downstream investigation stage and demonstrates the importance of managing the system load.

Section 6 discusses our approaches to alleviate endogeneity concerns and demonstrates the robustness of our main results. Section 7 details the roadmap for this ongoing project and enumerates our next steps.

2. Literature Review

This work is relevant to two main streams of literature: the child welfare system and human-AI teaming.

Child welfare systems. The child welfare system in the United States is comprised of a network of services and policies aimed at protecting children from maltreatment such as abuse and neglect. Recent literature discusses its structure while highlighting the challenges and areas that could be improved (Slaugh 2024).

The U.S. child welfare system is decentralized, with each state having its own system, though they all operate under federal laws and guidelines. Foster care, a critical component of child welfare, involves removing children from homes where they are at substantial risk of maltreatment and providing them with substitute care. Other essential aspects of the child welfare system in the US include foster care and adoption (Olberg et al. 2021).

One significant challenge is the overrepresentation of children of color, particularly Black children, in the child welfare system. Literature on racial disproportionality and disparities sheds light on the systemic and societal factors contributing to this issue and calls for targeted reforms (Doe and Clark 2020). Additionally, aging out of foster care without adequate support leads to higher risks of homelessness and unemployment among these young adults (Wilson 2019).

Existing literature underscores the complexity of the U.S. child welfare system, the myriad challenges it faces, and policy-making. However, research focusing on the operational improvement for the upstream services (e.g., call screening and investigation) in child welfare organizations is sparse (Slaugh 2024). This paper fills this gap by conducting a detailed empirical study of call screening at a child welfare organization in a US county.

Human-AI teaming. Many US states have implemented algorithms to assist child welfare operations (Saxena et al. 2020). Child welfare operations adopted these algorithms to reduce costs and, ideally, improve operational efficiency, equity, and service quality. Cheng et al. (2022) showed that screening decisions are more equitable when combining the strengths of AI algorithms and humans' clinical judgement. Fogliato et al. (2022) showed that call workers adjust their behaviors after the deployment of the AI tool. Call workers are capable of integrating complementary AI recommendations with their own judgement—Evidence show that they are less likely to adhere to erroneous AI recommendations. However, existing work on human-AI collaboration in a child welfare context does not consider vital operational factors such as workload and capacity.

In different context, Snyder et al. (2022) conducted a behavioral study that investigates humans' algorithm understanding and reliance under different pressure levels. The authors implemented laboratory experiments for a large-scale personalized recommendation context. Results show that greater time pressure increases human reliance on algorithms in general. Considering heterogeneous algorithm performances, humans rely more on superior algorithms as their ability to discern algorithm performance also improves under high load. To our best knowledge, we are the first to study human-AI teaming under various workload conditions utilizing a real-world datasets.

3. Empirical Setting and Data Description

This section presents the research setting, dataset and summary statistics, and data processing. We describe the key variables, dependent variables, control variables, outcome labels in detail.

3.1. Research setting

We focus our analyses on the screening stage of CWO operations. Specifically, we examine when human screeners' decisions align with or deviate from AI recommendations, and how call screening decisions are influenced by the workload. To study these research questions, we conduct empirical analyses on the referral level using the general form shown in Equations (1) and (2):

$$\text{Screening Decision} \sim \text{Workload} + \text{Controls} \tag{1}$$

$$\text{Deviation from AI} \sim \text{Worlload} + \text{Controls} \tag{2}$$

In particular, we would like to see how workload influence screening decisions and deviation from AI. Below we introduce the data and preprocessing (Section 3.2), the workload variables (Section 3.3), the dependent variables (Section 3.4), the outcome labels (Section 3.5), and controls (Section 3.6).

3.2. Data description

Our collaborating CWO granted us access to their private datasets, which are stored on a secure remote server hosted by a large research institution in the US. The datasets contain referral-level data from January 1, 2017 to November 11, 2022. For each incoming child maltreatment report, the datasets record the following information: the number of children involved and the child(ren)'s ID(s) in the system, demographic information about the household, zipcode, allegation type(s), abuse or neglect type(s), and the reporter's relationship to the household. From an operational standpoint, the system also records the date and time of receiving the report, the call screener's ID, and the supervisor's ID. The call screener also assesses the safety concerns and risks associated with each referral, and they document their individually assessed initial risk evaluations in three categories: High risk, medium risk, and low risk. Call screeners run the PRM tool after entering information about the child maltreatment report, and the system automatically outputs and stores

the PRM score. Based on the PRM score, risk protocols, and risk evaluation (which sometimes involves a field investigation), the screener makes a screening recommendation and enters it into the database. A supervisor then authorizes a final screening decision, which may differ from a screener's recommendation, after reviewing all the referral information mentioned above.

Besides referral information, our datasets also include investigation and case details. Recall that screened-in referrals are investigated and often assigned to another caseworker at a regional office. We have an assignment table that includes the caseworker IDs and times of investigation assignments. Investigation outcomes are summarized in service decisions: "accept for service," which will be followed by opening a case and providing services by the CWO and community partners, or "do not accept for service," closing the investigation without providing services. While our analyses focus on the screening stage and discuss screening decisions' impact on the investigation stage, our datasets contain additional details regarding the services provided to cases and case assignments.

Table 1 describes the summary statistics of relevant variables in the datasets.

Table 1 Summary statistics

	N	Mean	Median	Max	Min	SD
Case load	2140	1898	1028	2129	1637	167.45
Investigation load	2140	1047	1156	1468	530	164.54
Call load	2140	41.93	1028	2129	1637	20.42
Number of screeners	2140	13.94	15	18	3	4.74
Calls per screener	2140	2.85	2.88	5.5	1	0.66
Investigation workforce	2140	227.7	230.8	275	177	19.28
Case workforce	2140	305.7	306.6	365.8	29	27.98
PRM score	64870	13.95	15	20	1	4.57
Number of children	64870	5.142	5	22	1	2.06
Min. child age	64870	6.32	5	88	0	5.31

3.3. Key variables

The key variables of our analyses are the workload at the intake office and across the entire organization. The workload at the intake office is described in the variables "referral load," and "calls per screener." "Referal load" describes the number of incoming calls/referrals that are received by the intake office. We also track the "number of screeners" that received at least one call on a given day; there are always more than one screener on duty within our time window. We obtain the variable "calls per screener" by diving "referal load" by "number of screeners". Similarly, based on the start and end dates of investigations and cases, we obtain the number of active "investigation load" and "case load."

3.4. Dependent variables

The dependent variables of our main analyses include the "screen-in" decisions by supervisors and "deviation", whether the final screening decisions differ from those implied by PRM scores. In the main analyses, we define "deviation" as the follows: If the PRM score is equal or above 15, but the referral is screened out, then we label the screening decision a deviation from PRM recommendation. Similarly, if the PRM score is less than 15, but the referral is screened in, then the screening decision also deviates from PRM. Otherwise, there is no deviation. Note that the definition of deviation is mainly used for our analyses of human behaviors; it is different from that of PRM high or risk protocols which are implemented by the CWO and are applicable to only a fraction of referrals. On average, 36.03% of the time the final screening decisions deviate from PRM at our collaborating CWO.

We also study call screeners' recommendations which are yet to be reviewed by their supervisors. Similarly to the deviation definition above, we define recommendation deviation as when AI recommendations differ from the screeners' judgement.

3.5. Referral outcome labels

Consistent with CWO's most recent documentation (Allegheny County DHS 2024), we define the outcome of referrals as follows:

- True positive: If a referral is screened in (i.e., positive), and the investigation concludes that it requires service, i.e., "service decision" is true, then the initial screening decision outcome is labeled as true positive or TP.
- False positive: If a referral is screened in (i.e., positive), and the investigation determines that no service is necessary, i.e., "service decision" is false, then the initial screening decision outcome is labeled as false positive or FP.
- True negative: If a referral is screened out (i.e., negative), and no child on the referral is removed from the household in the next 90 days, then the initial screening decision outcome is labeled as true negative or TN.
- False negative: If a referral is screened in (i.e., positive), and at least one child on the referral is removed from the household in the next 90 days, then the initial screening decision outcome is false negative or FN.

The ultimate goal of the CWO is to provide services to families in need; this means the CWO aims to have a low FN rate. Meanwhile, the CWO strives to reduce FP rate, the reasons are twofold: First, investigating unsubstantiated and unfounded claims might cause unnecessary hassles and even harm to families. Second, the actual workload per caseworker at CWOs often exceeds the ideal levels; CWOs are therefore incentivized to reduce the number of investigations that do not lead to opening cases.

The CWO must strike a delicate balance at the screening stage: If screening in all referrals, the FN rate will be zero, but this is likely infeasible due to the capacity and budget constraint, and will increase FP rate which means higher level of unnecessary inconvenience or even interference. If screening out more referrals, the likelihood of having a FN outcome increases. The PRM tool is designed to help making informed screening decisions by providing the percentile catogories of predicted OOH placement likelihood. As a result of deploying PRM and other accumulative effort, the CWO has maintained a remarkable FN rate of 2.28%. (It is possible that the initial screen-out decision for a FN referral is actually accurate; as many things can change in 90 days.) However, the FP rate is 18.74%, which is higher than our collaborating CWO's expectation. There is an incentive within the CWO to reduce the FP rate while maintaining the low FN rate.

3.6. Control variables

Our CWO datasets enable us to use a comprehensive set of control variables in our empirical analyses. We describe our control variables below. The summary statistics for these control variables are presented in Table 1.

Demographics: We control for the demographics of the reporter, child(ren), and perpetrator associated with referrals. If a child is suspected of being maltreated, then entire household which the child belongs to is included in the referral. A household includes the victim(s) and all children with the same mother, as well as other adults living with the mother or children. All victims, children, reporters, and perpetrators are given unique IDs and their demographics data and interactions with the organization are recorded into CWO's database. Specifically, we account for the number of child(ren) associated with the household, the child(ren)'s age(s), race(s), household zipcode(s), the reporter's relationship with the victim(s).

Allegations and risks: For each referral, we control for the allegation type(s) and the PRM score, which describes the risk of at least one child being removed from the household in the next two years. We also flag families that are linked to an active case with the CWO with the variable "active family." About 16.13% of all referrals are linked to active families. These control variables help account for the nature of allegations and overall risks.

Organization workload and workforce: To better capture the organization's workforce and capacity, we include the "number of screeners" as a control. Similarly, based on how many workers are actively working on investigations and cases, we track the "investigation workforce" and "case workforce" on daily levels and include them as controls.

Time and holidays: We control for the year, month, and day of the week for our main analyses. We also present additional robustness checks in which we additionally account for weekends, winter and summer vacations, and holidays.

4. Empirical Investigation on Workload's Impact in Call-Screening

This section discusses the main analyses of workload effects on call screening decisions. Section 4.1 describes the model specification, and Section 4.2 presents the main regression results and their implications on human-AI teaming in multi-stage operations.

4.1. Model specification

In our primary empirical analyses, we investigate the effects of workload levels—both during the call-screening phase and across the entire system—on call-screening decisions and the collaboration between screeners and PRM. These analyses focus on GPS referrals, where call screeners and their supervisors have decision-making discretion. (Recall that CPS referrals are mandated to be investigated.) Our main regression specifications for the referral-level analyses are as follows: Let Y_j denote the supervisor-approved call screening decision for referral j arriving at time t (1 for screened-in and 0 for screened-out).

$$\text{logit}(Y_j) = \alpha + \beta_1 \text{Inv\_Load}_t + \beta_2 \text{System\_Load}_t + \text{Referral\_Controls} + \text{Time\_Controls} + \text{Screener\_FE} + \epsilon_j, \tag{3}$$

Let Z_j represent the deviation from the PRM suggested screening decision (1 for deviation and 0 for alignment) .

$$\begin{aligned}\text{logit}(Z_j) = \kappa + \gamma_1 \text{Inv\_Load}_t + \gamma_2 \text{Inv\_Load}_t^2 + \gamma_3 \text{System\_Load}_t + \gamma_4 \text{System\_Load}_t^2 \\ + \text{Referral\_Controls}_j + \text{Time\_Controls}_t + \text{Screener\_FE}_i + \epsilon_j,\end{aligned} \tag{4}$$

We incorporate comprehensive controls across referral characteristics, time, and screener dimensions; see details in Table 2.

4.2. Main regression results

Table 2 presents the main findings. Columns (1) and (3) illustrate the effect of increased workload on the probability of call screeners recommending and supervisors deciding to screen in a referral. The results suggest that humans account for system load and incoming call volume when determining screening outcomes. Specifically, we notice a decreased probability of screen-ins as the workload intensifies. Take results from the supervisor's decision as an example.

The column on the right presents the workload's impact on the probability of call screeners and supervisors deviating from risk protocols based on PRM scores (an AI model's output) and child ages. The results suggest a U-shape relationship between deviation and workload. In particular, human agents are more likely to diverge from the AI recommendations when the system workload is either high or low, rather than moderate. This might seem counter-intuitive at first. Yet a closer look reveals that while the AI tool does not adjust for varying workloads, human agents seem to factor in the workload when making screening decisions. Specifically, they lean toward admitting more low-risk cases when the system load is low and rejecting more high-risk cases when it is high. This behavior has the potential to enhance overall system performance, as will be discussed in Section 5.

Table 2 Workload's impact in call-screening

	Screen-In	Deviation
Key Variables:		
Investigation load	-0.004^{**} (0.002)	-0.0004 (0.008)
Investigation load (sq)		0.00002 (0.00004)
Case load	-0.028^{***} (0.003)	-0.128^{**} (0.050)
Case load (sq)		0.0003^{***} (0.0001)
Controls:		
Investigation workforce	0.012^{***} (0.004)	-0.018^{***} (0.004)
Case workforce	0.003 (0.002)	0.007^{***} (0.002)
Referral load	-0.003^{**} (0.001)	0.001 (0.001)
Number of call screeners	0.016^{***} (0.006)	0.002 (0.006)
PRM score	0.123^{***} (0.003)	0.046^{***} (0.003)
Number of children	0.057^{***} (0.007)	-0.010 (0.006)
Child min. age	-0.027^{***} (0.003)	-0.004^{*} (0.003)
Active family	2.444^{***} (0.048)	-1.073^{***} (0.037)
Race	YES	YES
Zip code	YES	YES
Allegation	YES	YES
Call screener ID	YES	YES
Reporter relationship	YES	YES
Holidays	YES	YES
Year, month, DoW	YES	YES
Observations	40,647	40,647

Note. "Investigation load" and "case load" are measured by the numbers of all active investigations and active cases (divided by 10) on a day, respectively. "Investigation load (sq)" and "caseload (sq)" are the squared terms of "investigation load" and "case load," respectively. "Investigation workforce" and "case workforce" describe the number of actively working investigation and case social workers in a day, respectively. "Referral load" is measured by the number of referrals received by the CWO on a day; "Number of call screeners" is the number of actively working call screeners on a day. "Number of children" is the number of children in the household in which maltreatment to at least one child is reported in the referral; "child min. age" is the minimum age of these children. "Active family" indicates whether the referral is associated with a family currently under investigation related to prior referrals or has been receiving ongoing services provided by the CWO. *p<0.1; **p<0.05; ***p<0.01

5. Empirical Evidence on the Importance of Managing Workload

So far, our results imply that call screeners incorporate both incoming and overall system workloads into their decision-making. This could potentially enhance the system's overall performance, underscoring the value of a collaborative decision-making process between humans and AI, rather than relying solely on AI. We delve further into this by examining why it is important to manage screen-in based on workload.

We first analyze the workload's impact on each individual investigation's duration. To study the impact of workload and capacity at the investigation stage, we use a linear model and control for the investigation's characteristics. Table 3 shows that the logged investigation duration is positively correlated with the logged investigation duration, but not with the investigation workforce.

Moreover, we analyze how increased downstream workload may adversely affect the quality of the subsequent investigation stage on a system level. We evaluate the efficiency of this phase using two system performance metrics: "investigation throughput," which refers to the number of closed

Table 3 Workload Impact on Investigation Duration from An Individual Investigation's Perspective

	Dependent variable:
	Logged investigation duration
Logged investigation load	0.088^{*} (0.050)
Investigation workforce	0.0003 (0.0004)
PRM score	-0.015^{***} (0.001)
Number of children	0.013^{***} (0.002)
Child min. age	0.006^{***} (0.001)
Race	YES
Zip code	YES
Allegation	YES
Call screener ID	YES
Reporter relationship	YES
Holidays	YES
Year, month, DoW	YES
Observations	21,794
R^2	0.042
Adjusted R^2	0.034

Note. "Logged investigation load" refers to the logged number of active investigations on a day averaged over the last thirty days. "Investigation workforce" refers to the number of active investigation workers on a day averaged over the last thirty days. We utilize a linear model to study how overall investigation load impacts each individual investigation's duration. *p<0.1; **p<0.05; ***p<0.01

Table 4 Workload's Impact on Investigation Efficiency From The System's Perspective

Dependent variable:	Investigation throughput	Investigation duration (logged)
	(1)	(2)
Investigation workforce	0.013^{***} (0.006)	-0.001 (0.001)
Investigation load	0.094^{***} (0.002)	
Investigation load (logged)		0.174^{***} (0.053)
Observations	1,195	1,195
R^2	0.894	0.471
Adjusted R^2	0.892	0.462

Note. "Investigation workforce" refers to the number of active investigation workers on a day averaged over the last thirty days. "Investigation load" refers to the number of active investigations on a day averaged over the last thirty days. In column (1), we utilize a linear model to compare estimation results against Little's Law, aiming to investigate how load impacts the extent of system congestion. In column (2), we implement a log transformation on both the investigation load and investigation duration to address the skewness present in both variables. Additional controls incorporated into the model include the day of the week, month, and year. *p<0.1; **p<0.05; ***p<0.01

investigations averaged over the last 30 days for each day recorded in our dataset; and "investigation duration," which refers to the duration of newly closed investigations averaged over the last 30 days for each day. (We employ a 30-day rolling average for these metrics to mitigate daily fluctuations.)

Table 4 shows that the throughput increases as the number of investigation workers or the number of screened-in referrals (i.e., incoming investigation volume) increases. However, for every

additional incoming investigation in a day, the average number of closed investigations on that day only increased by 0.094 (significantly lower than 1), suggesting an increased number of investigations in the system (i.e., congestion) and increased investigation load. Column (2) shows that an increased incoming investigation volume is often associated with a longer investigation duration. Interestingly, increase investigation workforce also lead to increased duration. There are several possible explanations: Increased workforce could mean that caseworkers can afford to be more thorough with each investigation without affecting the throughput. Also, an increased workforce allows caseworkers to work on those less urgent investigations with timelines that are not as tight. (Nevertheless, Table 3 shows that increased workload increases the duration of individual investigation completion, controlling for investigation characteristics.) Moreover, when there is 1% increase in the logged investigation load, the 7-day average investigation duration sees an increase by 0.12%.

6. Robustness Analysis

In our primary empirical analyses, as presented in Equation 3, it is important to address concerns related to endogeneity. Below, we discuss the approaches taken to alleviate potential endogeneity concerns associated with the call load.

First, in Section 4, we intentionally employ the shift-level workload (i.e., the daily average for all call-screeners and caseworkers) as our primary measure for the "call load," "investigation load," and "case load," instead of the workload for individual call-screeners and caseworkers. This is because individual workload might be subject to potential endogeneous assignment rules and practices. For instance, call screeners or investigation caseworkers could be assigned to certain types of referrals or investigations: Some might specialize in high-complexity (high-risk) reports while others focus on low-risk ones. If call screenes for low-risk referrals consistently handle more cases due to their simpler nature, we could see a trend where increased individual workloads correlate with decreased screen-in rates. However, this does not imply that a system-wide increase in referrals leads to lower screen-in rates. To avoid this endogeneity concern arising from assignments, the shift-level load is chosen to serve as a more reliable measure.

Nevertheless, even under the shift-level measure, there could be other endogeneity challenges. For example, a public awareness campaign or a traumatic event in the community may lead to increased number of child abuse/neglect calls. This would increase the overall shift-level load, but on average these calls might be of lower risks in nature.

To further alleviate such concerns, we consider using matching and weighting methods for call-screening, investigation, and case workload. This is an ongoing effort and preliminary results are promising. In addition, we also run the main analyses by controlling or excluding weekends, summer and winter vacations, as well as holidays. Results show that our main findings are robust.

7. Conclusion and Future Directions

This work studies the collaboration between human agents and the PRM in call screening decision-making. Our findings reveal that human agents are more likely to deviate from AI recommendations when faced with high or low workloads. Interestingly, while the PRM does not adjust for varying workloads, human agents appear to consider workload when making screening decisions, resulting in a U-shaped relationship between deviation and workload. We show in Section 5 that an increased investigation load might delay the completion of each individual investigation. These insights suggest that human workers effectively complement the AI's recommendations by incorporating important operational considerations such as maintaining a sustainable workload.

We have conducted extensive robustness checks which demonstrate that our main findings on the workload effect and the U-shape deviation patterns are strongly robust. We are exploring causal relationships between workload and screening decisions as well as human-AI collaborations through matching and weighting.

We aim to build upon our findings to offer recommendations to enhance the collaboration between human workers and AI by (i) incorporating measures of workload, (ii) proposing strategies to effectively improve screening decisions as well as the screening-investigation workflow, thus (iii) maintaining a sustainable system load. We provide evidence supporting that, by incorporating workload, call screeners can improve system welfare. Specifically, we have found that effectively managing workload enhances operational efficiency in the subsequent investigation process (recall Table 2). We enrich this analysis by considering load-aware screening protocols and how they might influence not only efficiency but also screening accuracy. Preliminary simulation results show that optimizing the risk protocol thresholds for default screen-ins and screen-outs can reduce the rate of screen-ins while maintaining and reducing false positive and false negative rates. This improvement requires minimal changes to the CWO's existing information system. Other promising avenues to effectively incorporate load in screening decisions include displaying downstream traffic, average workload per caseworker, investigation load, and average case durations to better inform call screeners of the actual business levels. Alternatively, the AI output could include the ranking of the risks and urgency of the incoming referral compared to incoming and existing referrals/investigations/cases in the system. However, these approaches require a greater amount of changes to the existing system and intricate analysis of caseworkers' mental models in making screening decisions.

Broadly speaking, our work contributes to the discussion on human-AI teaming in high-stakes decision-making within organizations with budget and capacity constraints. Like the PRM used in our partner organization, many deployed AI tools are programmed for a prediction task and cannot identify or address important operational constraints. We provide evidence showing that human

workers can complement AI tools by incorporating operational considerations (e.g., managing a sustainable workload). This finding again emphasizes the significance of effective human-AI collaboration; it also points to directions to enhance AI tools designed for use in high-stakes contexts.

References

Allegheny County DHS (2024) Allegheny family screening tool. `https://www.alleghenycounty.us/Services/Human-Services-DHS/News-and-Events/Accomplishments-and-Innovations/Allegheny-Family-Screening-Tool`, accessed: 2024-03-19.

CDC (2023) Fast facts: Preventing child abuse and neglect. `https://www.cdc.gov/violenceprevention/childabuseandneglect/fastfact.html`, accessed: 2023-06-13.

Cheng HF, Stapleton L, Kawakami A, Sivaraman V, Cheng Y, Qing D, Perer A, Holstein K, Wu ZS, Zhu H (2022) How child welfare workers reduce racial disparities in algorithmic decisions. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 1–22.

Doe C, Clark D (2020) Racial disproportionality and disparities in the child welfare system. Journal of Social Issues 76(4):765–788.

Fogliato R, De-Arteaga M, Chouldechova A (2022) A case for humans-in-the-loop: Decisions in the presence of misestimated algorithmic scores. Available at SSRN .

Michigan HHS (2023) Michigan's department of health and human services, children's protective services investigation process. `https://www.cdc.gov/violenceprevention/childabuseandneglect/fastfact.html`, accessed: 2023-06-19.

Olberg N, Dierks L, Seuken S, Slaugh VW, Ünver MU (2021) Search and matching for adoption from foster care. arXiv preprint arXiv:2103.10145 .

Pennsylvania DHS (2023a) Child protective services laws. `https://www.dhs.pa.gov/KeepKidsSafe/About/Pages/CPS-Laws.aspx`, accessed: 2023-06-19.

Pennsylvania DHS (2023b) Predictive risk modeling in child welfare in allegheny county. `https://www.alleghenycounty.us/Human-Services/News-Events/Accomplishments/Allegheny-Family-Screening-Tool.aspx`, accessed: 2023-10-12.

Pennsylvania DHS (2024) Mandated reporters: Frequently asked questions. `chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.dhs.pa.gov/KeepKidsSafe/Clearances/Documents/FAQ_Mandated%20Reporter.pdf`, accessed: 2023-06-19.

Saxena D, Badillo-Urquiola K, Wisniewski PJ, Guha S (2020) A human-centered review of algorithms used within the us child welfare system. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–15.

Slaugh V (2024) Berenguer G, Sohoni M, eds., Nonprofit Operations and Supply Chain Management: Theory and Practice (Springer Nature).

Snyder C, Keppler S, Leider S (2022) Algorithm reliance under pressure: The effect of customer load on service workers. Available at SSRN 4066823 .

Wilson F (2019) Aging out of foster care: A national comparison of homelessness and unemployment risks. Youth Studies Quarterly 38(1):22–34.