Human-Artificial Intelligence Teaming and Effects of System Load on Child Welfare Screening

Yanhan (Savannah) Tang[1], Zhaohui (Zoey) Jiang[1], Alan Scheller-Wolf[1], Justine Galbraith[2], and Lindsey Lacey[2]

[1] *Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA*
[2] *Allegheny County Department of Human Services, Pittsburgh, PA*

October 13, 2023

Abstract

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.

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; in 2018 the total lifetime economic burden of child abuse and neglect was about $592 billion. 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, 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 abuse 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). Many US states (e.g., Michigan, Pennsylvania) established laws that mandate CPS referrals to be investigated [Michigan HHS, 2023, Pennsylvania Department of Human Services, 2023a], while GPS referrals are only investigated by an assigned case worker 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.

1.2 Operations in a child welfare organization

We partner with a child welfare organization (CWO) within a county in North America. 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 is opened, and a social worker is assigned to intervene, offer protection, provide service, and work up a long-term solution for the child/children/family involved.

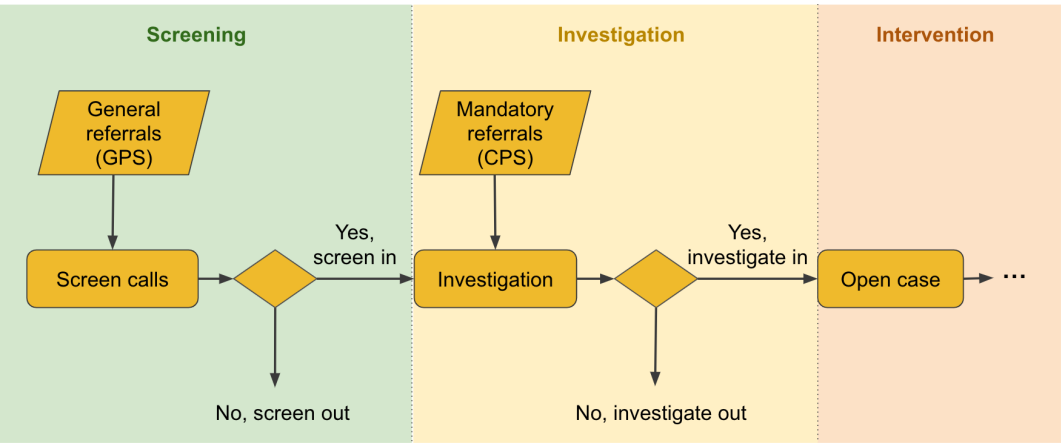

Figure 1: CWO workflow diagram.

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 higher probability of OOH placement within two years [Pennsylvania Department of Human Services, 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.

Figure 2 illustrates the workflow of the screening stage. When a call comes in, a call screener answers the call while recording information following the established protocols. Some referrals with specific characteristics fall into the *high-risk protocol* and, therefore, are encouraged to be screened in. In contrast, other referrals fall into the *low-risk protocol* and are encouraged to be screened out. Many other referrals fall into neither high-risk nor low-risk 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. Based on their assessment and the PRM risk score, the screener recommends screening in or screening out the referral. All screeners' recommendations go through their supervisors, and based on the PRM scores and screener recommendation (and often some discussions between the screeners and supervisors), as well as all relevant data, the supervisors make the final screening decisions.

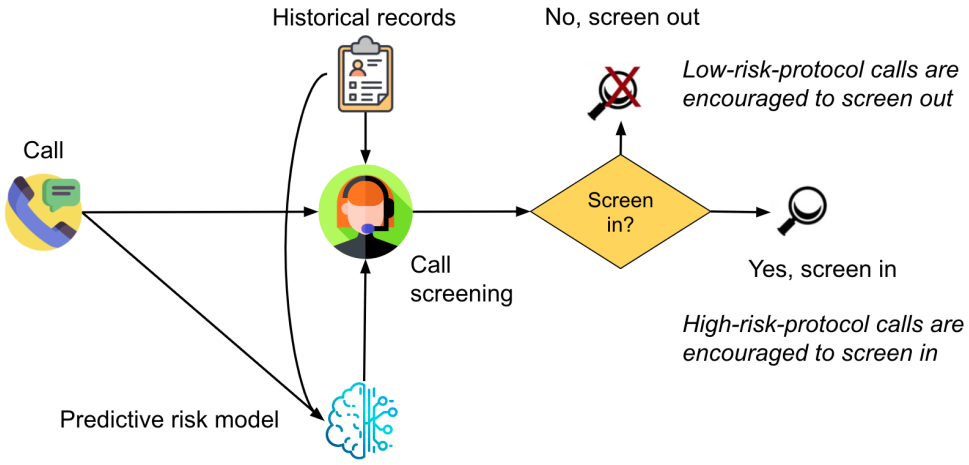

Figure 2: Workflow diagram: the screening step.

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. Our study explicitly focuses on the role of system load in human-AI collaboration in a high-intensity service environment with capacity constraints. 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 2 presents our main analysis results. Section 3 explores the impact of workload in the downstream investigation stage and demonstrates the importance of managing the system load. Section 4 discusses our approaches to alleviate endogeneity concerns. Section 5 details the roadmap for this ongoing project and enumerates our next steps.

2 Empirical Investigation on Workload's Impact in Call-Screening

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 have decision-making discretion. Our main regression specifications for the referral-level analyses are as follows:

$$\text{logit}(Y_j) = \alpha + \beta_1 \text{Call\_Load}_t + \beta_2 \text{System\_Load}_t + \text{Case\_Controls} + \text{Time\_Controls} + \text{Screener\_FE} + \epsilon_j, \tag{1}$$

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

where Y_j represents the decision for referral j arriving at time t (1 for screened-in and 0 for screened-out), and Z_j indicates the deviation from the PRM suggestion (1 for deviation and 0 for alignment)[1]. For Y_j and Z_j, we consider both the call worker's recommendations and their supervisor's final decisions. We incorporate comprehensive controls across referral, time, and screener dimensions; see details in Table 1.

Table 1 presents the main findings. Columns (1) and (3) illustrate the effect of increased workload on the probability of call-workers recommending and supervisors deciding to screen-in a case. 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 workload intensifies. Take results from the supervisor's decision as an example. For a unit increase in system load (or call load), we estimate a decrease of -0.93% (-2.66%) in the screen-in probability (calculated using the baseline screen-in probability of 47.6%).

Columns (2) and (4) present the workload's impact on the probability of call-workers and supervisors deviating from PRM suggestion. 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 lean toward rejecting more high-risk cases when the load is high. This behavior has the potential to enhance overall system performance, as will be discussed in Section 3.

[1]In consultation with our research partner CWO, referrals with PRM scores above certain thresholds are recommended by PRM to be screened in, while those with low PRM scores are suggested to be screened out. A referral j is considered a deviation $Z_j = 1$ when it is recommended to be screened-in (or out) based on PRM but is decided to be screened-out (or in) by the screener. Calls with moderate scores, between the two thresholds, can be screened in or out without being considered a deviation.

Table 1: Workload's Impact in Call-Screening

	Call-worker Recommendation		Supervisor Decision	
Dependent variables:	Screen-In	Deviation	Screen-In	Deviation
	(1)	(2)	(3)	(4)
Key Variables:				
System Load	-0.003^{**} (0.001)	-0.073^{***} (0.017)	-0.016^{***} (0.001)	-0.050^{***} (0.017)
System Load (squared)		0.0002^{***} (0.00005)		0.0001^{***} (0.00005)
Call Load	-0.011 (0.018)	-0.211^{**} (0.091)	-0.045^{**} (0.019)	-0.165^{*} (0.091)
Call Load (squared)		0.038^{**} (0.015)		0.035^{**} (0.015)
Controls:				
PRM score	0.102^{***} (0.002)	0.043^{***} (0.002)	0.061^{***} (0.002)	-0.085^{***} (0.002)
Risk protocol (=High)	0.725^{***} (0.085)	-0.579^{***} (0.095)	0.615^{***} (0.094)	-0.483^{***} (0.101)
Risk protocol (=Low)	-0.904^{***} (0.107)	-1.045^{***} (0.115)	-1.369^{***} (0.100)	-1.472^{***} (0.107)
No. child involved	0.064^{***} (0.005)	-0.010^{**} (0.004)	0.070^{***} (0.005)	-0.015^{***} (0.005)
Is active family (=True)	-1.473^{***} (0.030)	0.728^{***} (0.027)	-2.230^{***} (0.043)	0.857^{***} (0.029)
Race controls	Yes	Yes	Yes	Yes
Referral zip code	Yes	Yes	Yes	Yes
Allegation/abuse types	Yes	Yes	Yes	Yes
Reporter relationship	Yes	Yes	Yes	Yes
Year, Month, DoW	Yes	Yes	Yes	Yes
Screener FE	Yes	Yes	Yes	Yes
Observations	64,870	64,870	64,870	64,870
Log Likelihood	−36,373.930	−40,505.780	−33,408.550	−40,141.830
Akaike Inf. Crit.	73,259.870	81,527.560	67,329.100	80,799.670

Note. "Call Load" is measured by the number of referrals per screener on a day; "System Load" is measured by the number of all active cases (divided by 10) on a day. "Is 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

3 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.

In particular, we analyze how increased downstream workload may adversely affect the quality of the subsequent investigation stage. We evaluate the efficiency of this phase using two metrics: "investigation throughput," which refers to the number of closed investigations averaged over the last 7 days for each day recorded in our dataset; and "investigation duration," which refers to the duration of newly closed investigations averaged over the last 7 days for each day. (To mitigate daily fluctuations, we employ a 7-day rolling average for these metrics.)

Table 2 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 0.198 (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. Moreover, when there is 1% increase in the investigation

Table 2: Workload's Impact on Investigation Efficiency

Dependent variable:	Investigation throughput	Investigation duration (logged)
	(1)	(2)
Investigation workforce	1.200^{***} (0.296)	-0.004^{***} (0.001)
Investigation load	0.198^{***} (0.033)	
Investigation load (logged)		1.593^{***} (0.118)
Observations	1,568	1,517
R^2	0.253	0.395
Adjusted R^2	0.241	0.385

Note. "Investigation workforce" refers to the number of active investigation workers on a day averaged over the last seven days. "Investigation load" refers to the number of active investigations on a day averaged over the last seven 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

load, the 7-day average investigation duration sees a significant increase by 1.6%.

4 Efforts in Alleviating Endogeneity Concerns

In our primary empirical analyses, as presented in Equations 1 and 2, it is important to address concerns related to endogeneity. One element of particular concern is the call load. While the system load may not be directly connected to the focal referral j under investigation and therefore may be less problematic, the call load requires careful scrutiny. The underlying reason is that there could be confounding variables influencing both the incoming call volume and their screening decisions, since both are directly associated with the focal cases. It is crucial to ensure that these external factors do not inadvertently bias our findings. Below, we discuss the approaches taken to alleviate potential endogeneity concerns associated with the call load.

First, in Section 2, we intentionally employ the shift-level workload (i.e., the daily average for all call-screeners) as our primary measure for the Call_Load, instead of the workload for individual call-screeners. This is because individual workload might be subject to potential endogeneous assignment rules and practices. For instance, call screeners could be assigned to certain types of referrals: Some might specialize in high-complexity (high-risk) reports while others focus on low-risk ones. If screeners 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 instrumental variables (IV) for call-screening workload. Specifically, we use the number of call-screeners scheduled on a shift as an IV: It influences the workload (more screeners typically mean less work for each individual screener), and is unlikely to be directly connected to the nature of the referrals. Our main results remain qualitatively unchanged.

Ideally, an IV in our context should influence the call-screener's decisions solely through its effect on workload, without being correlated with the nature of the referral. In line with this, we intend to introduce an additional IV: The difference between the planned and actual number of screeners during a shift. Such differences, often arising from exogenous factors like unplanned vacations or sudden illnesses, is highly unlikely to be associated with the characteristics of the referrals (but will lead to changes in workload). Currently, we are actively working on this analysis.

5 Future Steps and Contributions

We are continuing to progress on the following aspects to further enhance the paper. First, as mentioned in Section 4, we are currently working on further strengthening the causal relationships between workload and screening decisions as well as human–AI collaborations.

Second, we plan on thoroughly exploring how, by incorporating workload, call screeners can improve system welfare. Specifically, we have found that managing workload is crucial to enhance operational efficiency in the subsequent investigation process (recall Table 2). We are working on enriching this analyses by considering how downstream workload may influence not only efficiency but also work quality. This will be examined by investigating its impact on the investigation error rate.

Lastly, 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; and (ii) proposing strategies to effectively improve screening decisions as well as the screening-investigation workflow; thus (iii) maintaining a sustainable system load.

Broadly speaking, our work contributes to the discussion on human-AI teaming in high-stakes decision-making. Like the PRM used in our partner organization, many deployed AI tools are programmed and cannot identify or address important operational constraints; 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

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

[Fogliato et al., 2022] Fogliato, R., De-Arteaga, M., and 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 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.

[Pennsylvania Department of Human Services, 2023a] Pennsylvania Department of Human Services (2023a). Child protective services laws. `https://www.dhs.pa.gov/KeepKidsSafe/About/Pages/CPS-Laws.aspx`. Accessed: 2023-06-19.

[Pennsylvania Department of Human Services, 2023b] Pennsylvania Department of Human Services (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.