Routinized Deference: Predictive Screening and Bureaucratic Capacity in Child Protective Services

Jiaen Wu

O'Neill School of Public and Environmental Affairs

Indiana University-Bloomington

jiaewu@iu.edu

Abstract: Predictive algorithms are increasingly used to support frontline decision-making in public administration, but evidence remains limited on how they affect bureaucratic capacity. This study examines county-level adoption of predictive screening tools in U.S. Child Protective Services using NCANDS data (2010–2023) and a stacked difference-in-differences design. Leveraging staggered adoption, I find no change in the total number of investigations opened and a sizable decline (about 15%) in the share of investigations that are substantiated. These results are inconsistent with operational capacity gains and more consistent with misclassification that erodes analytical capacity. These findings are consistent with a mechanism of routinized deference: under workload and accountability pressures, algorithmic recommendations can become procedurally defensible defaults. Overall, this study argues that predictive screening is not capacity-enhancing by default, as routinized deference can undermine the professional judgment central to bureaucratic work.

Keywords: routinized deference, predictive algorithms, bureaucratic capacity, algorithmic fairness, child protective services

Bureaucratic capacity—the human, technical, informational, and organizational resources that public agencies hold and can actually deploy in frontline work—is a central and persistent challenge in public administration (Ingraham et al., 2003; Bertelli and Lynn, 2006; Bello-Gomez, 2020). Frontline agencies often operate under significant resource constraints and high workloads, forcing street-level bureaucrats to develop discretionary shortcuts that can compromise the quality and equity of public service delivery (Lipsky, 1980). Enhancing this capacity, particularly the analytical ability to make professional judgments and the operational ability to process cases efficiently, remains a primary goal for public managers and policymakers (Drolc and Keiser, 2021).

The integration of predictive algorithms represents a promising technological solution to these capacity challenges. From local police departments forecasting crime to federal agencies detecting fraud, algorithmic tools are increasingly embedded in the core functions of government, often with the stated promise of enhancing bureaucratic capacity by increasing efficiency, standardization, and consistency (Eubanks, 2018; Meijer and Wessels, 2019; Mergel et al., 2023). In other words, by processing vast amounts of data, these predictive tools are expected to augment analytical capacity by improving decision quality and expand operational capacity by standardizing procedures and increasing throughput (Wu et al., 2015; Bullock et al., 2020).

Despite their rapid adoption, robust empirical evidence on how these algorithms actually affect bureaucratic capacity in practice remains limited. While some research suggests predictive algorithms can enhance bureaucratic performance (Kleinberg et al., 2018; Angelova et al., 2023), other studies reveal patterns of unintended human-algorithm interactions and institutional constraints that may amplify existing inefficiencies and erode capacity (Alon-Barkat and

Busuioc, 2023; Snow, 2021; Moynihan et al., 2025). These conflicting findings raise a crucial puzzle: why might a technology designed to augment bureaucratic capacity paradoxically undermine it?

While existing explanations for such unintended outcomes largely locate the problem in individual cognition—automation bias and complacency in how single decision-makers process algorithmic advice (Parasuraman and Manzey, 2010; Alon-Barkat and Busuioc, 2023)—less attention has been given to how organizations turn these individual tendencies into stable patterns of practice. Hence, this paper develops a theoretical framework centered on how the institutional context of street-level bureaucracy transforms algorithmic advice in practice. When workload and accountability pressures mount, algorithmic recommendations may become procedurally defensible defaults: easier to follow than to override, even when they are imperfect. I call this process routinized deference. The concept helps explain how tools designed to augment professional judgment may instead redirect analytical effort toward verification and exception management, such that they erode the very capacity they were meant to expand.

To empirically answer the question of how algorithmic governance influences bureaucratic capacity, I focus on predictive algorithms for detecting and preventing child abuse and maltreatment. In the U.S., some Child Protective Services (CPS) agencies have implemented predictive algorithms to support decision-making in the screening process on each referral to mitigate the workload for social workers (Gillingham, 2019). Hence, this paper uses the data from National Child Abuse and Neglect Data System (NCANDS) to estimate the impact of the implementation of predictive algorithm on the rate of substantiating reported cases of abuse or neglect.

Utilizing the year of algorithm implementation as a quasi-treatment within a stacked difference-in-differences framework, I find that adoption does not increase the overall number of investigations but is followed by a significant decline in substantiation rates. The concentration of the decline among high-yield professional referrals, together with rising re-referrals after unsubstantiated investigations, strengthens the interpretation that adoption reduced analytical capacity. To probe the behavioral mechanism indirectly, I also examine whether investigation decisions become more aligned with the pseudo-model risk score after adoption. This alignment strengthens only after repeated use, consistent with deference that consolidates through practice rather than arriving with the tool. Under shocks to the accountability environment—unexpected increases in state maltreatment fatalities—the alignment persists even as decisions in comparison states become less tied to underlying risk, providing suggestive evidence that algorithmic recommendations may become procedurally defensible defaults. Race-stratified analyses show no detectable new white/non-white divergence, a pattern consistent with parity by stagnation rather than an equity gain. Together, these results suggest that far from enhancing capacity, current algorithmic implementations foster routinized deference that gradually erodes the analytical core of bureaucratic work.

Algorithmic Governance and Bureaucratic Capacity

Recent advances in artificial intelligence (AI), particularly predictive algorithms and machine learning-based decision tools, have been posited as solutions to mitigate the cognitive and operational constraints (Eubanks, 2018; Meijer and Wessels, 2019; Mergel et al., 2023). This perspective is not merely academic. It is explicitly echoed in U.S. government policy. Across multiple administrations, Executive Orders (e.g., EO 13960, EO 14110, EO 14179) and OMB memoranda direct federal agencies to leverage AI to foster innovation, enhance government

operations, and improve the efficiency of public service delivery (Congressional Research Services, 2025). In other words, recent predictive algorithm tools are increasingly promoted as decision aids intended to support bureaucratic decision-making, offering AI-generated recommendations that aim to assist street-level bureaucrats in making more consistent and unbiased decisions (Ludwig et al., 2024).

Under this view, we can consider bureaucratic capacity as a stock of resources, such as the human, technical, informational, and organizational assets an agency holds (Ingraham et al., 2003; Bertelli and Lynn 2006; Christensen and Gazley, 2008; Bello-Gomez, 2020). Accordingly, when an agency adopts AI tools, those tools are expected to enhance its capacity along two dimensions: analytical capacity, or the ability to interpret complex information and detect patterns in high-dimensional data in order to support sound decision-making; and operational capacity, or the ability to process cases efficiently, consistently, and at scale (Wu et al., 2015). First, AI systems may strengthen analytical capacity by rapidly synthesizing large volumes of data and detecting patterns beyond human cognitive reach. For example, Kleinberg et al. (2018) conducted a quasi-experimental study showing that predictive algorithms can improve judges' analytical capacity on whether to detain or release individuals, ultimately reducing crime rates. Similarly, Angelova et al. (2023) examined judges' decisions to override algorithmic recommendations on jailing or release decisions and found that only approximately 10% of judges demonstrate superior accuracy and fairness compared to the algorithms. Second, operational capacity can be expanded as algorithms reduce processing times and introduce standardized decision-making procedures across administrative settings, thereby improving throughput and consistency (Bullock et al., 2020). For example, in policing and risk-based workplace safety inspections, predictive algorithms enhance operational capacity and standardize

previously inconsistent decision-making processes, thereby improving overall agency throughput (Mohler et al., 2015; Johnson et al., 2023).

However, considering capacity as a stock of resources builds the optimistic conclusion into the concept itself: adopting a tool is, by definition, an addition to the stock, so adoption can only be considered as a gain. The difficulty is that a resource an agency possesses is not the same as a resource it can deploy. What governs case-level decision quality is not the assets an agency holds but the share left free for individualized judgment once routine demand is met (Nohria and Gulati, 1997; Elston and Zhang, 2025). This distinction is important for studying AI adoption, given that AI shapes decisions through its use under workload rather than through its presence in the agency's inventory (Williams, 2021). In other words, the same tool can add capacity when it offloads routine cases and frees workers for complex ones, or drain it when the time it frees is reabsorbed into verifying scores, managing exceptions, and clearing queues. Whether AI adoption increases or decreases capacity—and whether it does so through the analytical or the operational dimension—is ultimately an empirical question. This distinction aligns with recent calls to evaluate AI not only by model performance, but by its downstream effects on organizational work, judgment, and capacity in practice (Hauser et al., 2025).

Empirical findings on algorithmic adoption are mixed. Some studies find gains in efficiency or consistency (Kleinberg et al., 2018; Angelova et al., 2023), while others find unintended consequences that undermine the capacities these tools were meant to expand (Alon-Barkat and Busuioc, 2023; Moynihan et al., 2025). A growing literature suggests that these divergent outcomes turn less on whether algorithms are present than on the ways in which decision-makers interact with algorithmic recommendations in practice (Kellogg et al., 2020; Waardenburg et al., 2022; Vaccaro et al., 2024). The dominant explanations for problematic interaction operate at the

individual level. Parasuraman and Manzey (2010) showed that decision-makers under high workloads monitor automated systems less attentively (automation complacency) and treat automated advice as a heuristic substitute for their own information search (automation bias). In public sector settings, such over-reliance is further patterned by decision-makers' priors: bureaucrats adhere to algorithmic advice most readily when it matches stereotypes of client groups (Alon-Barkat and Busuioc, 2023) or confirms their prior professional judgment (Selten et al., 2023). These accounts locate the problem of human-AI interactions in individual cognition shifts, implying that over-reliance on AI reflects changes in individual decision-making processes. Yet street-level bureaucrats are not isolated information processors. They make decisions within organizations that assign caseloads, audit dispositions, and allocate blame when decisions go wrong (Lipsky, 1980; Hood, 2011). However, less attention has been given to how this organizational context transforms individual tendencies toward over-reliance into organizational patterns of practice.

I therefore propose the concept of "routinized deference" to describe the process by which reliance on algorithmic recommendations hardens from an individual cognitive shortcut into an organizationally expected mode of operation: a default decision routine sustained not by trust in the tool's accuracy but by the procedural defensibility of following it and the personal cost of departing from it. I develop the concept in two steps: the driver that sustains deference, and the process through which it consolidates.

The first feature concerns what sustains deference: the asymmetry of accountability inside public organizations. Blame for failure weighs more heavily on front-line bureaucrats than credit for success (Weaver, 1986; Nielsen and Moynihan, 2017). Under this accountability pressure, decision-makers therefore move to the option that is easiest to justify (Lerner and Tetlock, 1999),

and the algorithmic score supplies such an option. More specifically, the organization's accountability structure embeds an asymmetry in the cost of following versus declining the algorithm's recommendations. A worker who follows the score and errs is procedurally covered: the decision conformed to the sanctioned instrument, and the presence of an algorithmic system further complicates how responsibility for the outcome is attributed (Busuioc, 2021; Alon-Barkat et al., 2025). A worker who overrides the score and errs, on the other hand, owns the error personally, and the override itself typically demands justification in the form of documentation, supervisory sign-off, and a defensible reason to depart from the number (Porter, 1995; Espeland and Sauder, 2007; Artinger et al., 2019). Deference, in this account, is less a cognitive failure than a rational adaptation to a blame-averse organizational environment (Hood, 2011).

The second feature concerns how deference consolidates through repetition in the front-line workflow. Organizational routines form as recurrent practice stabilizes into procedural memory (Cohen and Bacdayan, 1994) and acquires taken-for-granted status (March and Simon, 1958; Nelson and Winter, 1982; Feldman and Pentland, 2003). Deference to the algorithmic score follows the same pattern. Early in adoption, workers treat the recommendation as one signal to weigh against their own reading of a case (Christin, 2017; Brayne and Christin, 2021) and override recommendations they judge to be wrong (De-Arteaga et al., 2020). With repetition and sustained accountability pressure, however, verification gives way to default. For example, in a police organization observed over three years, Waardenburg et al. (2022) show that stable reliance on algorithmic outputs was not present at deployment but emerged gradually through organizational practice. Studies of hospital settings find the same erosion: clinicians under time pressure gradually come to rely on AI decision support and, after a period of routine use, exhibit

"deskilling," with lower detection rates when the AI is absent (Rinta-Kahila et al., 2023; Budzyń et al., 2025).

Routinized deference is therefore expected to erode the analytical capacity of front-line bureaucrats. As defaulting to the algorithm becomes standard practice, the cognitive effort once devoted to independent diagnosis is crowded out; analytical work migrates from diagnosis to exception management, and capacity is reallocated away from individualized appraisal even as some operational metrics improve. In the ideal version of human-AI interactions, front-line bureaucrats retain their own judgment precisely because even a well-designed screening tool will misrank some cases. When workers read a case independently, the two imperfect readings—the worker's and the model's—check one another, and a misleading score can be caught before it becomes a disposition. Routinized deference, however, disables that check: the decision collapses onto the model's reading alone, and the tool's errors pass through into dispositions uncorrected. In this way, adoption can increase the stock of analytical resources an agency holds while shrinking the analytical capacity it deploys at the case level. Such arguments lead to the following hypothesis:

H1: The adoption of predictive algorithms will reduce bureaucratic analytical capacity.

Second, routinized deference reshapes the distributive pattern of bureaucratic capacity, raising crucial questions about equity. A significant body of literature highlights the risk of algorithmic discrimination (Eubanks, 2018). When inputs are incomplete, mismeasured, or unrepresentative, outputs are more likely to systematically misclassify some cases and overlook others (Arnold et al., 2021). Basu (2021) formalizes how measurement error in outcomes or predictors can generate algorithmic discrimination even when designers seek fairness, and Basu et al. (2024) show in health contexts that group-differential measurement error can yield

discriminatory predictions despite race-neutral specifications. Similarly, other studies show that homogenized scoring and proxy-based objectives can create correlated errors that systematically disadvantage specific groups, thereby widening disparities (Obermeyer et al. 2019).

Conversely, some scholars argue that algorithms can mitigate existing human biases, thereby promoting fairness. From an equity perspective, standardized discretion, a key characteristic of artificial discretion, helps mitigate the potential threat of bureaucratic discretion abuse (Bullock, 2019; Bullock et al., 2020). In other words, algorithms offer consistency and data-driven accuracy that can curb the disparate treatment resulting from human prejudices. For example, Kleinberg et al. (2018) found that artificial discretion could reduce racial bias in pretrial detentions by eliminating some inconsistent and overly harsh decisions made by human judges. Similarly, Bartlett et al. (2022) found that AI-based credit scoring was more consistent and less biased than credit assessments by human loan officers.

This paper reframes the debate by proposing a third possibility rooted in routinized deference. Because deference is sustained by an asymmetry in accountability that workers share, rather than by any single worker's cognition, adoption does not merely change individual decisions but homogenizes them. Cases that once passed through different workers, each applying their own reading, now pass through a single scoring logic, a condition Kleinberg and Raghavan (2021) call algorithmic monoculture. Prior research on human–AI interaction anticipates new biases rooted in workers' priors: bureaucrats adhere most readily to advice that matches stereotypes of client groups (Alon-Barkat and Busuioc, 2023) or confirms their prior professional judgment (Selten et al., 2023), such that adoption reproduces or even widens pre-existing patterns of discretionary variation. The expected distributive consequence of routinized deference, by contrast, is neither targeted harm nor an equity gain but what I term parity by

stagnation: groups remain comparable to one another not because the agency serves them equitably but because its capacity to serve them erodes at a common rate.

And this uniform misclassification will burden the system from two directions. On one side, false positives, incorrectly flagging an individual for intervention or inclusion, often lead to the misallocation of agency resources, whether through wasteful investigations or improper benefit distribution. On the other, false negatives, failing to flag an individual who warrants intervention or inclusion, can represent a core failure of the agency's mission. Such errors impose significant administrative burdens on clients, who must navigate complex appeals or seek redress. This, in turn, generates a reactive workload for the agency, as staff time is consumed by responding to these complaints and managing the crises that may result from the initial failure. Moynihan et al. (2025) characterize such environments as governed by "clawback logics," in which automation channels small score biases into recovery actions and documentation cycles that absorb staff time without improving underlying judgments.

This feedback loop, where the agency must manage the fallout from both error types, means operational capacity fails to improve while analytical capacity erodes. The burden falls not necessarily on a specific demographic group, but on the system's overall efficiency. This mechanism yields a second hypothesis:

> H2: The adoption of predictive algorithms will lead to a uniform erosion of bureaucratic capacity without creating new between-group disparities.

The effect of predictive algorithms on child abuse and maltreatment

This paper focuses on predictive algorithms used in detecting and preventing child abuse and maltreatment. In the United States, child abuse and maltreatment pose a serious problem.

According to CDC statistics, approximately 1 in 7 children experienced child abuse in 2023. Additionally, the Children's Bureau estimated that around 600,000 children were victims of maltreatment, and 1,820 children died from neglect and abuse in 2021 (Administration for Children and Families, 2023).

CPS agencies are generally housed within state departments of human services and carry out casework through local or county offices.[1] Initially, upon receiving a referral, intake workers decide whether to screen in or screen out.[2] Common screen-out reasons include allegations unrelated to abuse or neglect, insufficient information, jurisdictional conflicts, or involvement of children over 18 years (Administration for Children and Families, 2022). Screened-in referrals move to investigation, where caseworkers assess whether maltreatment occurred. This decision-making process, both with and without algorithmic support, is a complex exercise in professional judgment. Caseworkers must synthesize a wide array of evidence, from tangible documents like medical and police reports to more subjective assessments based on interviews with the child, caregivers, and other sources. They must also weigh immediate safety threats against longer-term risks, and balance child vulnerabilities against caregiver protective capacities (Font and Maguire-Jack, 2022). In most states, these factors are weighed against a "preponderance of the evidence" standard, meaning a report is substantiated if it is determined that maltreatment was "more likely than not" to have occurred. Substantiated cases may then receive post-response services such as foster care (Administration for Children and Families, 2022).

[1] 41 states run CPS through a centralized state system, 9 delegate primary authority to counties, and 2 use hybrid systems.

[2] At CPS intake, referrals are screened in when they meet statutory/policy criteria for a CPS response—becoming "reports" that receive investigation or an alternative response; referrals that do not meet these criteria are screened out.

Each year, CPS agencies handle roughly 4.28 million screened-in referrals nationwide, placing immense pressure on frontline staff. While a comprehensive national figure for caseloads per worker is unavailable, data from reporting states reveal the extreme and uneven nature of these workloads. For instance, staffing levels vary widely across reporting states: Wisconsin's screeners handled about 16 cases per worker in 2022, whereas Missouri's averaged about 1,754; the weighted national ratio across reporting states is about 273 cases per worker (median ≈ 364) (Administration for Children and Families, 2022).[3] Excessive caseloads assigned to individual child protection workers have been linked to a higher probability of burnout (Fryer Jr and Miyoshi, 1989; Conrad and Kellar-Guenther, 2006) and compassion fatigue (Conrad and Kellar-Guenther, 2006; Sprang et al., 2011), ultimately leading to decreased performance (Leake et al., 2017; McFadden et al., 2018). To relieve workload and curb unwarranted variation in discretionary decisions, some state and county CPS agencies have begun adopting predictive risk models for the screening stage. These tools are not issued by a single national authority; instead, they are developed locally or purchased from vendors, leading to a patchwork of models across the country (Gillingham, 2019). For instance, many jurisdictions now contract with Eckerd Connects to use its Rapid Safety Feedback (RSF) predictive-analytics model during screening. RSF was first developed in Florida and applies historical child-welfare and related administrative data to flag referrals that resemble past cases with severe harm so supervisors can give those referrals extra scrutiny (Commission to Eliminate Child Abuse and Neglect Fatalities, 2016).

More generally, available descriptions of child-welfare predictive analytics characterize these systems as decision-support tools that draw on historical child-welfare records and, in some

[3] Author's calculation using screened-in referral and staffing data from the 2022 Child Maltreatment report. The national weighted ratio and median are calculated using the subset of states for which both data points were available.

jurisdictions, linked administrative data to estimate risk at the front end of the CPS process (ASPE, 2017; Vaithianathan et al., 2019). Predictive models process referral information and administrative histories and translate them into outputs such as risk scores. Depending on the jurisdiction, these outputs may inform screen-in decisions and prioritization, though they are generally presented as aids to professional judgment rather than as fully automated substitutes for caseworker decision-making (Gillingham, 2019; Cheng and Chouldechova, 2022). While the specific inputs, display formats, and workflow rules vary across tools, the common change is the insertion of predictive risk information into CPS screening decisions (Inman and Adams, 2026).

Furthermore, child protection is a classical blame-avoidance environment. Child maltreatment fatalities routinely trigger media scrutiny, legislative inquiry, and disciplinary or even criminal action against individual caseworkers, and a substantial literature documents the defensive practice this environment produces (Regehr et al., 2010; Gainsborough, 2010; Chenot, 2011; Jagannathan and Camasso, 2017). For example, research on the actuarial risk-assessment instruments (checklist-based tools such as structured decision-making frameworks) found that frontline workers used them less as diagnostic aids than as devices for justifying decisions to supervisors and external reviewers (Gillingham and Humphreys, 2010). Predictive screening tools inherit this function, and their deployment protocols formalize it. While workflow rules vary across jurisdictions, publicly documented implementations describe design features that attach procedural weight to the score, such as supervisory approval requirements for departures from the recommended action (e.g., Vaithianathan et al., 2019). Some qualitative studies of workers using predictive algorithmic risk systems describe the burden of documenting and justifying decisions that depart from the displayed score (Cheng and Chouldechova, 2022; Kawakami et al., 2022).

While the primary aim of predictive algorithms is to alleviate the burden on child protection workers, some research suggests that these screening tools may actually exacerbate existing problems. For instance, Gillingham (2019) conducted a study on screening tools used in Philadelphia and Illinois, raising concerns about the accuracy of the predictive algorithm results. They argue that inaccuracies in the predictive algorithm could lead to overestimation of maltreatment cases for thousands of children, thereby increasing the burden on child welfare workers. Conversely, Marshall and English (2000) and Schwartz et al. (2017) have expressed optimism regarding the accuracy and effectiveness of screening tools in detecting child abuse. However, most recent studies on the effectiveness of these tools have focused on only a few cases within specific agencies. This ongoing debate on the tools' fundamental effectiveness mirrors the broader theoretical idea that whether AI adoption can expand or erode the bureaucratic capacity.

Moreover, another significant thread in the literature addresses the social equity of predictive algorithms in child protection screening, specifically examining whether their predictive algorithms disproportionately flag historically over-surveilled groups. For example, Cheng et al. (2022) analyzed data from Allegheny County and found that predictive algorithms indeed generate racial disparities, but intake workers play a role in mitigating this effect. Conversely, Rittenhouse et al. (2022), utilizing quasi-experimental methods on Allegheny County data, concluded that predictive algorithms help to mitigate racial disparities. This unresolved debate in the child welfare context mirrors the broader theoretical tension regarding algorithmic fairness and directly motivates this paper's investigation into the distributive patterns of algorithmic impacts.

Data and Method

This paper utilized the National Child Abuse and Neglect Data System (NCANDS) from 2010 to 2023, which collected case-level data from all 50 states. NCANDS achieved this by compiling electronic records tailored to each child for reports of suspected child abuse and neglect handled by CPS (OASH, 2024). The data encompass child demographic information, types of maltreatment, levels of maltreatment disposition, information on child and caregiver risk factors, and records of post-investigation services.

NCANDS records final determinations in seven categories—substantiated, indicated/reason to suspect, alternative response victim, alternative response nonvictim, unsubstantiated, unsubstantiated due to intentionally false reporting, and closed–no finding—where substantiated denotes that CPS concluded the legal threshold for maltreatment was met. Building on these dispositions, I operationalize two facets of capacity at the investigation stage: operational capacity as the volume of investigations initiated, and analytical capacity as the substantiation rate conditional on investigation (the share of screened-in reports ultimately substantiated). To analyze policy-relevant variation, I aggregate case-level NCANDS records to the county–year level, producing a panel that tracks investigation volume and substantiation rates across counties and over time.

Because prior work documents variation in CPS contact across demographic groups, equity implications remain an active area of inquiry in child-welfare risk modeling (Wildeman and Emanuel, 2014; Maguire-Jack et al., 2021; Stapleton et al., 2022). Accordingly, I also construct race-stratified versions of both outcomes—investigation volume and substantiation rate—for white and non-white children. I rely on a white/non-white comparison because more disaggregated racial and ethnic categories produce sparse county-year cells once the data are stratified by event time and treatment cohort. While this aggregation improves statistical

stability, it also masks potentially important heterogeneity among Black, Hispanic, Native American, Asian, and other non-white children.

This study uses the report by Samant et al. (2021) to identify the implementation year of the predictive algorithm in each county. The authors compiled data from multiple publicly available sources, including news articles, predictive tool developers' websites, and official reports by state child welfare agencies, and supplemented this with a questionnaire sent to state agencies to verify the year of implementation. Samant et al. (2021) classified adoption into five stages: In use, Consulting, Explored, Dropped, and None found. For this analysis, I define treatment counties as those classified as "in use," meaning the predictive tool was actively used either on a pilot or regular basis. I exclude counties labeled as "explored" because this category indicates that the tool was not implemented or that no confirmable discontinuation could be found, making implementation status uncertain. These counties are excluded from both the treatment and control groups to avoid misclassifying counties with uncertain treatment status. Counties categorized as Consulting (where development, adoption, or partnerships were being explored) and those with None found (no mention of the tool in official or media sources) are included in the control group, since there is no evidence of actual tool use. For the counties categorized as Dropped, the post-discontinuation county-years are excluded from the treated exposure window rather than retained as treated or recoded as controls.

Although the tools coded as "in use" vary in vendor and input data, they share a common organizational function: each introduces predictive risk scoring into the CPS screening process to guide screen-in decisions (ASPE, 2017). I therefore treat adoption as exposure to a class of predictive-screening interventions rather than to a single standardized algorithm. My estimates capture the average effect of this common risk-scoring function across heterogeneous local

implementations. Table 1 organizes the years when counties started to implement this new technology.[4]

<Table 1 Here>

My empirical strategy examines changes in the annual investigation-stage outcomes—the number of investigations initiated and the substantiation rate—in counties where the algorithm was implemented relative to those where it was not, both before and after the implementation of the predictive algorithm for the screening process. Because two-way fixed-effects (TWFE) estimators confound treatment effects that arrive in different years, I adopt Wing et al.'s (2024) stacked difference-in-differences approach to estimate the results. Specifically, I carve the data into a series of balanced sub-experiments (one for every adoption year a) and then append the sub-experiment panels row-wise into one stacked dataset.

For each county i, adoption cohort a, and event time $e = t - a$, let

$$y_{iae} = \sum_{k=-k_{pre}}^{K_{post}} \beta_k \, Algorithm_{ia} \, I(e = k) + \gamma_i + \delta_e + \varepsilon_{iae}$$

As previously mentioned, my dependent variable y_{iae} is the investigation-stage outcomes for each county i, in event year e, within the sub-experiment whose focal adoption year is a. The variable $Algorithm_{ia}$ equals 1 if the county i belongs to the treatment arm of sub-experiment a, and 0 otherwise. The variable $I(e = k)$ is an indicator for event year k. The baseline year is one year before the implementation (event year (k) = -1), and the β_k represents the change in the

[4] Samant et al. (2021) only collected data until 2021. After 2021, I verified county use via searches of major news databases, state/county procurement records, CPS agency minutes/reports, and vendor announcements.

investigation-stage outcomes in the treatment county at year k relative to the control county, compared to the control year. γ_i and δ_e is county and event time fixed effects.

Each observation is weighted by the product of an inverse-probability weight (IPTW) that balances covariates within a sub-experiment and a corrective "stack" weight that restores balance across sub-experiments. For the inverse-probability weights, the propensity score is estimated using a logistic regression that includes an extensive set of county-level variables measured in 2010—the first year observed for every county and well before any algorithm adoptions. Specifically, the model includes baseline health outcomes (premature-death rate; share reporting poor or fair health; average days of poor physical and mental health; low-birth-weight rate; teen-birth rate), insurance status (share of uninsured adults), education and labor-market indicators (high-school-graduation rate, unemployment rate), economic wellbeing (children in poverty, income-inequality ratio), food access (share with reasonable access to healthy foods), and family structure (share of single-parent households). By balancing counties on this comprehensive set of pre-algorithm characteristics, the IPTW step helps ensure that treated and control counties are comparable on factors that could simultaneously affect both the decision to adopt the predictive tool and subsequent substantiation outcomes. The weighting values are calculated as:

$$w_{ia} = IPTW_{ia} \times Q_{ia}$$

$$IPTW_{ia} = Algorithm_{ia} + (1 - Algorithm_{ia}) \frac{\hat{\pi}(X_{ia})}{1 - \hat{\pi}(X_{ia})}$$

$$Q_{ia} = \begin{cases} 1, & if\ Algorithm_{ia} = 1 \\ \dfrac{N_a^T / N_\Omega^T}{N_a^C / N_\Omega^C}, & if\ Algorithm_{ia} = 0 \end{cases}$$

, where N_a^T (N_a^C) is the treated (control) sample size in sub-experiment a and N_Ω^T (N_Ω^C) is the total across all admissible adoption year Ω_k. Hence, the previous regression is estimated by weighted least squares with weights w_{ia} and standard errors clustered at the county level.

Overall, each coefficient β_k summarizes the estimated difference between adopting and comparison counties k years before or after adoption, measured relative to the year immediately before adoption ($k = -1$). For example, β_2 captures whether adopting counties changed more or less than comparison counties between the baseline year and two years after adoption. Because treatment timing varies across counties, each β_k aggregates the cohort-specific difference-in-differences contrasts available at that event time under the stacked DiD weights. The pre-treatment coefficients ($k < 0$) serve as lead estimates for assessing differential pre-trends, and the post-treatment coefficients ($k \geq 0$) trace the dynamic impact of the predictive algorithm on investigation-stage outcomes.

Results

Following my identification strategy, I estimate a stacked difference-in-differences model on the full sample to assess how algorithmic tools affect bureaucratic capacity in child protection screening. Firstly, I focus on analytical capacity, using the substantiation rate as the outcome. As shown in Figure 1, the pre-treatment coefficients are generally indistinguishable from zero. The coefficient at t= -4 is marginally significant (-0.022; p<0.10), but its small magnitude and isolated nature suggest no systematic trend. Consistent with this interpretation, a joint Wald test of the pre-treatment coefficients is not statistically significant (p-value = 0.18). Overall, the pattern of the pre-treatment coefficients supports the parallel-trends assumption. In the post-treatment period, Figure 1 shows no effect in the first two years, but a statistically significant

decline thereafter. The average post-treatment effect is −0.027, representing a substantive 15.2% reduction from the baseline substantiation rate of 0.178.

<Figure 1 Here>

While this result indicates a meaningful shift in frontline outcomes, it does not, on its own, confirm that algorithmic tools undermine bureaucratic analytical capacity. Instead, the decline could be explained by two different mechanisms. First, the misclassification channel holds that imperfect model accuracy leads to errors in prioritization. This could manifest as either diverting caseworker effort toward lower-yield referrals (false positives) or failing to flag high-risk referrals for necessary scrutiny (false negatives), either of which would be expected to reduce the overall substantiation rate. In other words, a misclassification channel would lower analytical capacity by diverting attention to low-yield referrals, which should appear as a decline in the substantiation rate without a corresponding rise in investigations and with longer investigation durations as exception management and corrective checks accumulate. Second, the efficiency channel suggests that automating routine screening tasks increases operational capacity to free up caseworker capacity, allowing them to initiate a greater number of investigations, including lower-yield cases, and thereby diluting the share of cases that meet the substantiation threshold. To adjudicate between these explanations, I next examine how the algorithm affects total investigation volume. If the decline stems from operational improvements, we should observe an increase in the number of investigations. If, however, no operational capacity improvements are observed, this would point toward analytical misclassification as the dominant mechanism.

In Figure 2, I regress the total number of investigations initiated per county on event-time indicators using the same stacked DiD design. All pre- and post-treatment coefficients are statistically insignificant (with only the $t = 1$ estimate marginally significant). In other words, this

result counters the efficiency-channel argument that the predictive algorithm increases operational capacity and allows case workers to investigate more cases. Figure 3 presents the event-study estimates for the average number of days per investigation. The pre-treatment coefficients remain tightly clustered around zero, with only one minor outlier that appears by chance rather than as part of any trend. The post-treatment coefficients are statistically significant. The mean post-treatment effect is 7.36 days, which, relative to a baseline of 52.01 days in event year -1, corresponds to an approximate 14 percent increase in investigation duration per case. This pattern suggests that instead of streamlining screening workflows, the algorithm is associated with longer case processing, further challenging the notion of operational improvement. This interpretation, however, rests on the key assumption that algorithm adoption does not coincide with other major shifts, such as changes in agency resources or front-end screening patterns. To formally test these alternative explanations, I conducted two supplementary analyses (see Appendix 1). The results confirm that neither the volume of screened-out referrals nor frontline staffing levels changed significantly after algorithm adoption.

<Figure 2 Here>

<Figure 3 Here>

To further unpack the drop in substantiation following algorithm introduction, I stratified reports by source yield. I define high-yield referrals as those from professional reporters (social services personnel, medical personnel, mental health staff, law enforcement, and education personnel) and low-yield referrals as those from informal sources (other relatives, friends or neighbors, alleged perpetrators, and anonymous reporters). Figure 4 plots the stacked DiD estimates for high-yield referrals. All pre-treatment coefficients are statistically indistinguishable from zero, while the post-treatment coefficients in event years +3 and +4 are sharply negative

and significant ($p < 0.05$), indicating a sustained decline in substantiation rates for high-yield cases. In contrast, Figure 5 shows that none of the post-treatment coefficients for low-yield referrals is significantly different from zero, and the substantiation rate among informal sources remains flat after the algorithm's application.

An efficiency-driven increase in capacity would predict a dilution of substantiation across all referral types, or even an uptick in low-yield investigations, yet I observe a decline only among high-yield cases. While this evidence does not identify mechanism on its own, it is more consistent with a misclassification channel in which professionally sourced reports are de-prioritized relative to their true risk. The algorithm systematically incorrectly ranks high-risk, professionally sourced reports as low-risk, causing caseworkers to under-investigate them (i.e., affording them routine rather than heightened investigative effort) and driving down the substantiation rate exclusively among high-yield referrals. In sum, the post-adoption decline in the substantiation rate, unchanged investigation counts, and longer case durations indicate declining analytical capacity (supporting H1) and no increase in operational capacity.

<Figure 4 Here>

<Figure 5 Here>

The previous evidence is consistent with a misclassification channel, but it cannot by itself rule out another interpretation: the decline in substantiation reflects improved triage rather than eroded capacity. If pre-adoption substantiation rates were inflated—if caseworkers historically over-substantiated professionally sourced referrals by weighting source credibility above case evidence—then a tool that redistributes effort away from these cases would be correcting a prior bias, and falling substantiation would signal improvement rather than eroded capacity. The high-

yield stratification weighs against this interpretation but does not fully rule it out. To adjudicate between the two, I examine a downstream outcome more proximate to child safety: whether children whose index investigation closed unsubstantiated are subsequently re-referred to CPS.

The two interpretations make opposing predictions on the re-referral rates. Under beneficial correction, cases newly diverted out of substantiation are ones the algorithm correctly identifies as low-risk; re-referral among this group should not rise. Under capacity erosion, the unsubstantiated pool comes to contain genuinely high-risk cases that the algorithm misranked as low-risk; re-referral among this group should rise. Hence, I construct a child-level measure—the share of children with an index unsubstantiated investigation who receive a new screened-in report within twelve months of the index report date—and estimate the stacked DiD design. Figure 6 presents the event-study estimates. The pre-treatment coefficients are clustered near zero, and the post-treatment coefficients turn positive and significant beginning in event year +2. The average treatment effect implies a 1.12 percentage point increase relative to a baseline rate of 15.84%, a 7.1% rise. The timing of this increase coincides with the substantiation decline: both outcomes remain flat for the first two post-adoption years and change only at event year +2. This pattern indicates that the substantiation decline reflects an erosion of analytical capacity rather than an improvement in triage.

<Figure 6 Here>

The evidence assembled so far—a decline in substantiation, no increase in investigations, longer case durations, a concentration of decline among professionally sourced referrals, and a rise in re-referral rate—indicates that algorithm adoption eroded analytical capacity through misclassification. But these results, on their own, cannot prove that the erosion of analytical

capacity is due to routinized deference. In other words, this raises the question of how frontline staff used the tool, and whether they deferred to an error-prone score rather than exercising the independent judgment that the tool happened to mislead.

There are two potential mechanisms through which predictive algorithms could cause capacity erosion, and both begin from the same premise: that the tool generates flawed rankings. They differ in what they assume about worker behavior. First, workers continue to exercise independent judgment: they weigh the score as one input among several, but outcomes deteriorate because that input is systematically misleading. Second, under the routinized deference mechanism developed above, workers increasingly default to the score in place of independent appraisal, so the tool's flawed rankings pass into dispositions uncorrected.

Two features of the evidence distinguish these mechanisms. First, the timing aligns more closely with routinized deference. If the tool were simply inaccurate and worker behavior did not change, its error profile would be fixed at deployment and would not evolve over time, and the effects should appear immediately. However, most of the capacity outcomes move after event year 2. The sustained break after year 2 is consistent with the possibility that workers still check the tool against their own judgment at the beginning, but over time, leaning on the score becomes the default, and judgment gives way. In other words, the two-year delay traces how deference to algorithmic recommendations consolidates into workers' routines and gradually displaces independent judgment. While the earlier increase in investigation duration might appear to cut against this interpretation, duration captures downstream investigation length rather than the brief screening decision where deference would save time. Where flawed rankings redirect attention toward the wrong cases, investigations may run longer even as the initial use of the score becomes more routine.

Second, to more directly test how workers interact with the algorithmic risk score, the ideal test would require the actual risk score for each case. However, the deployed model's specifications are not public. Instead, I approximate it with a pseudo-algorithm trained on historical data using similar modeling procedures (Allegheny County, 2019). Hence, I replicate a predictive screening model using NCANDS data from 2000 to 2009. Specifically, I train a LASSO logistic regression to predict substantiation outcomes based on detailed child-, caregiver-, and perpetrator-level variables. After sorting each child's referrals chronologically and retaining only the first report per child, the final analytic sample includes over 18 million unique investigations. I use reports from 2000 to 2007 (approximately 14.6 million cases) as a training set and reserve the 2008 to 2009 data (around 4 million cases) as a strict holdout validation cohort, ensuring that no child appears in both sets (see Appendix 3 for the detailed training procedure). When applied to a held-out 2008–2009 validation cohort, the model achieves an area under the curve (AUC) of 0.931, indicating strong predictive performance.

Using the pseudo risk score for each case, I estimate the gradient of investigation decisions on the case-level risk score for each county-year, within the stacked DiD design. This gradient serves as a proxy for how strongly a high-risk score inclines workers toward investigating a case. Figure 7 shows that the gradient steepens significantly after adoption (average post-treatment effect = 0.054). The increase builds over time: it is present but modest in the first post-adoption years and rises sharply from year 2, mirroring the delayed trajectory of the substantiation and re-referral outcomes. This pattern indicates that workers' decisions came to track algorithmic recommendations increasingly closely over the post-adoption period, which points to routinized deference as the more likely mechanism.

<Figure 7 Here>

This analysis relies on a pseudo risk model rather than the deployed risk model, and it is important to acknowledge that the two differ. Most notably, the pseudo-algorithm omits cross-system inputs, such as criminal-justice or housing records, that some deployed tools incorporate. Two considerations limit the consequences of this gap for my conclusions. First, the analysis does not require the pseudo-score to match the deployed score exactly, only to rank cases by risk in broadly the same order. Even if the pseudo-score is a noisy approximation of what workers actually saw, this noise would weaken the measured gradient rather than inflate it. That I nonetheless find a significant gradient implies that, with the true score, workers' reliance would appear at least as strong. Second, identification rests on how the gradient changes after adoption, not on its cross-sectional level: because the pseudo-score is generated by a single model, any systematic gap between it and the deployed score is differenced out and does not account for the post-adoption trend. Hence, the pseudo-score serves as an informative proxy for the changing interaction between workers and the algorithm, even if it does not reproduce the deployed tool exactly.

Figure 7 establishes that frontline decisions came to track the algorithmic ranking increasingly closely after adoption. What the alignment pattern cannot reveal, however, is why workers increasingly follow the score. Two mechanisms could produce this trajectory. The first is individual and cognitive: the automation complacency documented in prior work (Parasuraman and Manzey, 2010), in which workers habituate to the tool as trust in its accuracy accumulates with repeated use or lean on it more heavily as attention is stretched under load. The second is routinized deference, in which deference is sustained by the asymmetry between following the score and bearing the personal cost of overriding it. The two mechanisms are difficult to separate directly, as administrative data contain no measures of override requirements or workers'

perceived blame exposure. I therefore exploit shocks to the accountability environment, on which the two accounts make competing predictions. Under automation complacency, if workers follow the score because accumulated experience has built trust in its accuracy, a salient system failure—a case the system did not prevent—should shake that trust and loosen alignment with the score. On the other hand, under routinized deference, if workers follow the score instead for the procedural protection, the same failure should have the opposite effect: by raising the personal stakes of an unjustifiable decision, it raises the value of a procedurally defensible decision rule. Where an adopted tool is in place, following the algorithm supplies such a rule, and workers should remain anchored to the risk score even as blame exposure rises.

I use state-level maltreatment fatalities as shocks to the accountability environment. Fatal child-maltreatment cases are highly salient failures that trigger media scrutiny, legislative attention, and organizational pressure extending well beyond the office in which a death occurred (Gainsborough, 2010; Chenot, 2011). I construct annual state fatality totals from the NCANDS and standardize each state's fatalities relative to its 2010—2012 baseline[5]. The main measure is the larger of the standardized fatality surprises observed during the prior two years, allowing the accountability environment created by a salient case to persist beyond the year of the death. The analysis returns to the stacked difference-in-differences two-way fixed-effects regression behind Figure 7, aggregated to the state level, and adds one term: the interaction between the fatality surprise and the average treatment effect variable. The fatality surprise coefficient captures how decisions respond to blame pressure in comparison states, and the interaction captures how differently adopting states respond after adoption.

[5] Because NCANDS records fatalities only at the state level, this analysis is conducted at the state level

Table 2 shows that comparison states and adopting states respond differently to these accountability shocks. Among comparison states, a one-standard-deviation increase in the fatality surprise is associated with a 0.0040 decline in the risk-score gradient (SE = 0.0017, p = 0.023): under blame pressure, protective decisions become less selective with respect to underlying risk. This pattern is consistent with the defensive dynamics in prior work: publicized fatalities trigger indiscriminate expansions of protective action at the system level (Gainsborough, 2010; Chenot, 2011) as well as defensive practice among workers under blame exposure (Jagannathan and Camasso, 2011; Jagannathan and Camasso, 2017), whose risk judgments are themselves unstable under stress (Regehr et al., 2010; LeBlanc et al., 2012). The post-adoption treatment effect interaction with fatality surprise is positive (0.0027, SE = 0.0013, p = 0.039): following a recent fatality surprise, decisions in adopting states remain more closely aligned with the risk score than decisions in comparison states. In other words, once an algorithm has become embedded in the workflow, its recommendation supplies a sanctioned and readily defensible focal point, and decisions remain anchored to the risk ranking as blame exposure rises. This asymmetry supports the notion that routinized deference is sustained not simply by confidence in predictive accuracy or an individual cognitive shift, but by the procedural protection afforded by following an organizationally authorized recommendation.

<Table 2 Here>

The preceding evidence supports H1, indicating a decline in analytical capacity driven by misclassification. Hypothesis 2 addresses the distributive pattern of this decline, positing a uniform erosion of capacity rather than the creation of new racial disparities. To test this hypothesis, I estimate heterogeneous effects by race. I first verify that predictive algorithm introduction did not alter the trend in the total number of investigations by race. Using the same

stacked DiD design, I estimate event-time coefficients for the count of investigations initiated separately for white and non-white children. As shown in Figure 8, none of the post-adoption coefficients deviate significantly from zero for either group, indicating that overall investigation volume remained unchanged across races following implementation.

I then re-estimate the stacked DiD model for substantiation rates, stratifying the sample by race. Figure 9 presents the event-study estimates: circles denote white children, and triangles denote non-white children. Throughout the post-treatment period, the confidence intervals for the two series overlap almost entirely, suggesting no persistent racial disparity in the algorithm's effect on substantiation. Importantly, the average post-treatment effect is negative for each group, consistent with uniform inefficiency (a race-constant erosion of capacity). A triple difference-in-differences specification (see Appendix 2), in which the treatment coefficients are interacted with a non-white indicator, yields only one marginally significant difference at event time = 3, which is not sustained in subsequent years. Based on these results, I do not detect new racial disparities in substantiation, and the post-adoption decline in analytical capacity appears uniform across groups.

<Figure 8 Here>

<Figure 9 Here>

The race-stratified stacked DiD shows no detectable divergence in either investigation counts or substantiation rates between white and nonwhite children. This parallel pattern in post-treatment outcomes does not, by itself, establish that predictive algorithm is race neutral. It suggests two interpretations that connect directly to the capacity framework. One possibility is that the algorithm exhibits uniform misclassification: rankings are similarly off across groups, so

substantiation rates decline for every group without producing new racial gaps. A second possibility is that the algorithm carries racial bias at the prediction stage, but frontline bureaucrats offset it in use. Caseworkers need to expend additional time and judgment to correct group-specific errors. In other words, under uniform misclassification, the model's predictive behavior should be similar across groups. Under bias offset, the model should exhibit race-specific differences in predictive performance, while investigation-stage outcomes remain parallel because caseworkers expend additional time and judgment to correct group-specific errors.

To adjudicate between these possibilities, the ideal test would evaluate the algorithm actually in use. As discussed before, the deployed model's specifications are not public, so I use the same pseudo-algorithm described above. With this pseudo-algorithm, I compute annual county-level risk-score variation, defined as the within-county standard deviation of predicted risk scores, for white and non-white children on the 2010 to 2023 NCANDS sample and estimate separate event-study analyses. The goal of this analysis is not to identify a causal effect on the scoring rule, which is immutable, but to diagnose any shifts in the underlying risk profile of referrals at the time of adoption. Figure 10 shows that differences in score variation between the two groups are not statistically significant and that both series exhibit nearly identical trends. Within this NCANDS-based approximation, the pseudo-model assigns risk scores to white and non-white cases in broadly similar ways.

<Figure 10 Here>

Extending the event-study analysis, I assess the pseudo LASSO-based risk model's discrimination by race using NCANDS data from 2010 to 2023. This analysis tests whether the model's predictive performance differs across racial groups. Separate ROC curves yield AUC

(white) = 0.927 and AUC (non-white) = 0.909, an absolute gap of 0.018. While DeLong's test indicates this difference is statistically significant ($p < 0.05$), both AUCs exceed 0.90, indicating only a small difference in predictive accuracy across racial groups. This small ΔAUC indicates the model discriminates very slightly better for white children than for non-white children, but both curves remain well above conventional adequacy thresholds (>0.9). Crucially, the near-parallel receiver operating characteristic (ROC) curves (see Figure 11) imply that the relative ranking of high- versus low-risk cases is almost identical across racial groups.

<Figure 11 Here>

The absence of post-adoption divergence between white and non-white children warrants caution. It is consistent with a regime in which a common scoring logic homogenizes discretion, generating correlated errors and lowering average decision quality even in the absence of shocks, a key risk of algorithmic monoculture (Kleinberg and Raghavan, 2021). Bovens and Zouridis (2002) also argue that the shift from street-level to system-level bureaucracy channels discretion into ICT-driven routines, standardizing decision criteria and shrinking the space for professional inference. In the CPS setting, such parity in trends coexists with degraded performance (slower processing and lower yield), indicating a distributionally uniform erosion of capacity rather than an equity gain.

Because these race-stratified diagnostics rely on the same pseudo-model as the gradient analysis, they carry the same caveat discussed above: the pseudo-model approximates but does not fully replicate the tools counties actually deployed. The main concern is that deployed tools incorporate cross-system inputs, such as criminal-justice records and housing records, that are likely more racially patterned than the data available in NCANDS, such that deployed scores could be more race-differential than my approximation. However, Figures 8 and 9 show no

detectable post-adoption divergence in investigation counts or substantiation rates between white and non-white children, and the gradient analysis in Figure 7 suggests that investigation decisions became increasingly aligned with risk-score rankings after adoption. If deployed scores were strongly race-differential, and if those differences were transmitted into frontline decisions without offsetting behavior by caseworkers, we would expect some corresponding divergence in investigation volume or substantiation.

Taken together, the stacked DiD and the LASSO-based analyses inform Hypothesis 2. The results show that adoption does not generate new racial disparities in investigation or substantiation, and the pseudo-model exhibits high predictive performance for both groups with only small AUC differences. At the same time, the substantiation rate declines at a similar pace across groups. The capacity loss therefore appears to operate along non-racial lines, most plausibly through mis-prioritization by reporter type that reduces yield without increasing throughput. In short, the evidence is consistent with uniform inefficiency rather than group-specific harm.

Discussion

Using a stacked difference-in-differences design, I find that the introduction of predictive algorithms does not alter the total number of investigations but leads to a significant decline in substantiation rates beginning in the third year after adoption. The average post-treatment decrease is 0.027 points, or 15.2 percent relative to the baseline rate of 0.178. Mechanism tests reveal that investigation volume remains unchanged while average investigation duration increases by 7.36 days (14 percent), suggesting no gain in operational capacity. The added time also does not appear to reflect more thorough casework: longer investigations did not translate into more accurate dispositions, as re-referrals among unsubstantiated cases rose. The decline in

substantiation is concentrated among high-yield referrals from professional reporters, whereas low-yield referrals remain stable; this offers evidence of a misclassification mechanism that reflects a failure in analytical capacity. The re-referral analysis weighs further against the interpretation that lower substantiation reflects beneficial correction. Where index investigations closed unsubstantiated, the children involved become more likely to receive a new screened-in report within twelve months of the index report, which suggests that some cases moved out of substantiation were not simply low-risk cases correctly triaged away.

The results—investigation counts do not rise, yields fall (especially among professionally sourced referrals), re-referrals increase among previously unsubstantiated cases, and investigations take longer—are inconsistent with efficiency gains. A more plausible interpretation is that these findings are consistent with routinized deference: under workload and accountability pressures, the algorithmic score becomes a procedurally defensible default, and it is easier for frontline staff to follow the recommendation than to spend scarce analytical time justifying departures from it. As this pattern repeats across the screening workflow, analytical effort may shift from independent appraisal toward verification and exception management. The risk-score gradient analysis lends further support to this interpretation, in that investigation decisions become increasingly aligned with risk-score rankings after adoption. Because this gradient steepens over time rather than appearing immediately at deployment, it is more consistent with a gradual routinization process than with a fixed algorithm-quality problem or individual cognitive shift. The fatality-shock analysis adds a further piece of suggestive evidence on what sustains this alignment. If workers followed the score because accumulated experience had built trust in its accuracy, a salient system failure should weaken that trust and loosen alignment with the score. Table 2 instead shows the opposite asymmetry: under recent fatality

surprises, decisions in comparison states become less selective with respect to underlying risk, while decisions in adopting states remain anchored to the risk-score ranking. This pattern is consistent with the accountability asymmetry at the core of routinized deference: procedural protection of following an organizationally sanctioned recommendation becomes more valuable precisely when blame exposure rises.

However, this evidence should be read as indirect. The data show changing outcome patterns, increasing score-alignment, and differential responses to accountability shocks, but they do not directly observe the behavioral and cognitive process through which caseworkers incorporate algorithmic recommendations into routine practice. As recommendations settle into background infrastructure, they function less as intelligence multipliers and more as queue-clearing rules of thumb, with the attendant risk of an algorithmic monoculture that standardizes yesterday's discretion (Kellogg et al., 2020; Waardenburg et al., 2022; Kleinberg and Raghavan, 2021). Over time, this may further diminish analytical engagement, turning the tool into an object of routinized deference rather than a capacity-enhancing aid.

In testing Hypothesis 2, I find no evidence from the stacked DiD analysis that the algorithm differentially affects substantiation rates across racial groups. Event-study estimates for white and non-white children show nearly identical post-treatment trends, and a triple-difference model reveals no sustained racial divergence. To probe whether this reflects a race-neutral tool or bias that is offset in use, I replicate a predictive model using historical data. Because the pseudo-model is built on NCANDS rather than on the tools counties actually deployed, these pseudo-model diagnostics should be read cautiously. With that caveat, the pseudo-model produces similar risk-score dispersion across white and non-white children, and the ROC curves show only minor differences in predictive performance. This evidence does not establish that deployed

tools were race-neutral. Rather, together with the stacked DiD results, it suggests that adoption did not produce detectable new white/non-white divergence in investigation-stage outcomes. The pattern is therefore consistent with what I call parity by stagnation, a distributive implication of routinized deference. Adoption appears to channel cases through a common scoring logic that regularizes discretion, while analytical capacity declines at a similar rate across the two aggregated racial groups. When the absence of detectable white/non-white divergence is read in this light, it should not be taken as an equity gain, but as evidence consistent with uniform capacity loss under routinized deference.

Conclusion

In this paper, I leverage staggered county-level adoption of predictive screening tools in child protective services to find that algorithm adoption is associated with a capacity contraction, not an expansion: investigation volumes hold steady, but substantiation drops and processing slows. Race-stratified analyses show these negative effects are not detectably concentrated in one racial group in the white/non-white comparison. I interpret these findings as consistent with routinized deference, an organizational process in which algorithmic recommendations become procedurally defensible defaults under workload and accountability pressure and gradually crowd out analytical capacity. This interpretation helps explain why a tool designed to increase efficiency may instead be associated with slower, lower-yield work centered on verification and exception management.

This study makes several key contributions to public administration theory. First, it refines our understanding of bureaucratic capacity in the age of algorithmic governance. Building on bounded rationality and street-level discretion research, this paper proposes the concept of

routinized deference to theorize how predictive tools can function less as intelligence multipliers and more as queue-clearing infrastructure, compressing the space for professional inference.

Second, the findings highlight a concerning long-term trajectory of capacity erosion driven by flawed human-AI collaboration. The results of slower, lower-yield work are consistent with a growing body of evidence showing that human-AI teams often underperform without the right institutional conditions (Sele and Chugunova, 2024; Vaccaro et al., 2024). The mechanism behind this underperformance lies less in individual cognitive shifts than in organizational routines and accountability asymmetries. When following the score is easier to justify than departing from it, repeated use can turn algorithmic advice into a default reference point. Over time, this behavioral pattern may contribute to deskilling, as over-reliance erodes the fine-grained professional judgment needed for complex cases. The ultimate risk is that organizations do not learn or improve, but merely use technology to standardize yesterday's discretion (Kellogg et al, 2020; Rahman, 2021; Waardenburg et al., 2022).

Finally, this paper enriches the debate on algorithmic fairness by providing evidence consistent with "parity by stagnation." The uniform distribution of negative effects across the two aggregated racial groups should not be read as an equity gain. Rather, it points to the possibility of an "algorithmic monoculture" in which a common scoring logic regularizes discretion and produces similar capacity losses across groups (Kleinberg and Raghavan, 2021), a loss that appears distributionally neutral in aggregate yet remains substantively harmful.

For practice, these findings suggest that capacity should be evaluated as an attribute of use, not merely of accuracy. If agencies are to employ predictive tools, institutional design must be aligned with the realities of street-level work: protect analytical time, create feedback and learning loops between outcomes and practice, and assess performance with work tempo and

yield alongside predictive metrics (Bovens and Zouridis, 2002; Levy et al., 2021; Grimmelikhuijsen and Meijer, 2022). These design choices shape whether algorithms function as decision aids that sustain learning, or as routines that crowd it out.

This study has several limitations that suggest avenues for future research. First, relying on administrative data, this analysis cannot directly observe the cognitive or behavioral processes of routinized deference. Future qualitative or ethnographic research could provide rich, on-the-ground evidence of how caseworkers interact with these tools. Second, this study documents an average pattern across heterogeneous local implementations, but variation likely exists across organizational contexts. A related limitation is that a binary treatment indicator cannot capture the partial or gradual implementation that can occur within treated counties. Future research could explore how factors like training or leadership might mitigate or exacerbate the tendency toward routinized deference. Finally, the pseudo-algorithm, while informative, is an approximation of the proprietary tools used in practice, and its performance may differ from the undisclosed models deployed by various agencies.

In sum, the promise of predictive algorithms as capacity-expanding is not self-executing. Capacity hinges less on the algorithmic recommendation itself than on the organizational conditions that govern its use. Absent those conditions, routinized deference may emerge or persist, and capacity may quietly erode.

Table 1. Predictive Algorithms Implementation

State	Policy Adoption Year	Policy Dropped Year
Alaska	2016	2018
Arkansas	2020	
California	2016	2018
Colorado	2017	
Connecticut	2016	2019
Florida	2013	
Illinois	2016	2017
Indiana	2018	
Louisiana	2018	2019
Maine	2016	
New Hampshire	2018	
New York	2017	
North Carolina	2016	
Ohio	2017	2018
Oklahoma	2016	
Oregon	2018	2022
Pennsylvania[6]	2016	
Tennessee	2017	
Wisconsin	2013	2018

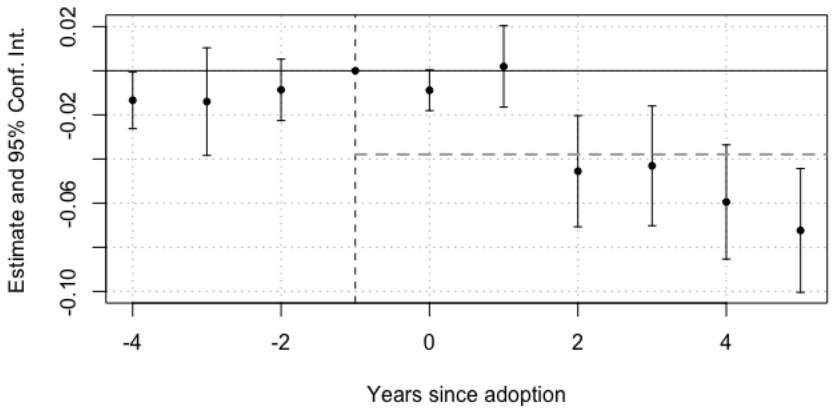
[6] Only Allegheny County adopted the predictive algorithms.

Table2. Stacked DiD Estimates of Fatality Shocks' Effect on the Risk-Score Gradient

	Risk-score slope
Fatality Surprise	-0.0040**
	(0.0017)
Fatality Surprise * Treated-post	0.0027**
	(0.0013)
Observations	1184
R2	0.771
Within R2	0.014

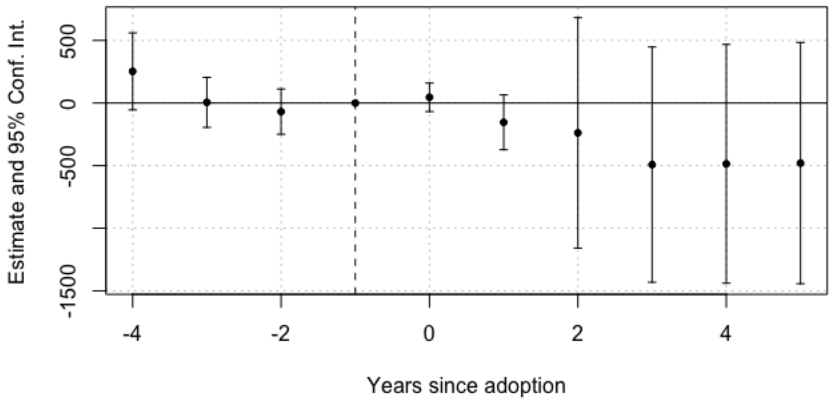
The estimation uses the same stacked difference-in-differences two-way fixed-effects regression as Figure 7, with the fatality-surprise interaction added. The dependent variable is the risk-score gradient underlying Figure 7, aggregated to the state–year–adoption-stack level. “Fatality surprise” is the standardized fatality surprise over the prior two years, standardized against each state’s 2010–2012 baseline. “Treated-post” is the average treatment effect variable as in Figure 7. The specification retains the adoption state fixed effects and event-time fixed effects. Standard errors are clustered by state. *p < 0.10, **p < 0.05.

Figure 1. Stacked DiD Estimates of Algorithm Adoption's Effect on Substantiation Rates



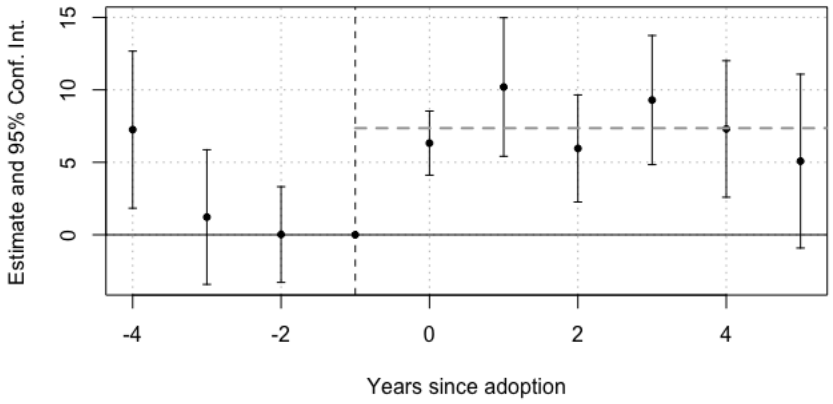
Note: Points report event-time estimates from the stacked difference-in-differences specification, and vertical bars show 95% confidence intervals. Event time is measured relative to the adoption year. The vertical dashed line marks the omitted reference period (k= -1). The solid horizontal line at zero indicates no estimated difference relative to that reference period. The gray horizontal dashed line denotes the average post-adoption effect across event years k=0 to k=5.

Figure 2. Stacked DiD Estimates of Algorithm Adoption's Effect on Total Investigations



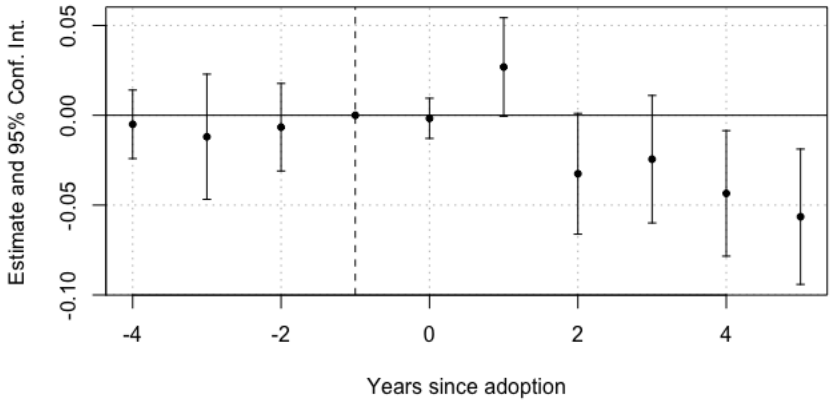
Note: Points report event-time estimates from the stacked difference-in-differences specification, and vertical bars show 95% confidence intervals. Event time is measured relative to the adoption year. The vertical dashed line marks the omitted reference period (k= -1). The solid horizontal line at zero indicates no estimated difference relative to that reference period.

Figure 3. Stacked DiD Estimates of Algorithm Adoption's Effect on Investigation Times



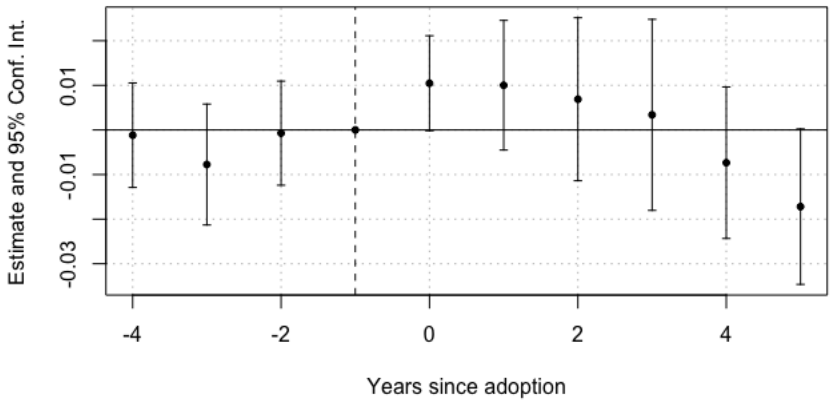
Note: Points report event-time estimates from the stacked difference-in-differences specification, and vertical bars show 95% confidence intervals. Event time is measured relative to the adoption year. The vertical dashed line marks the omitted reference period (k= -1). The solid horizontal line at zero indicates no estimated difference relative to that reference period. The gray horizontal dashed line denotes the average post-adoption effect across event years k=0 to k=5.

Figure 4. Stacked DiD Estimates of Algorithm Adoption's Effect on Substantiation Rates for High-Yield (Professional) Referrals



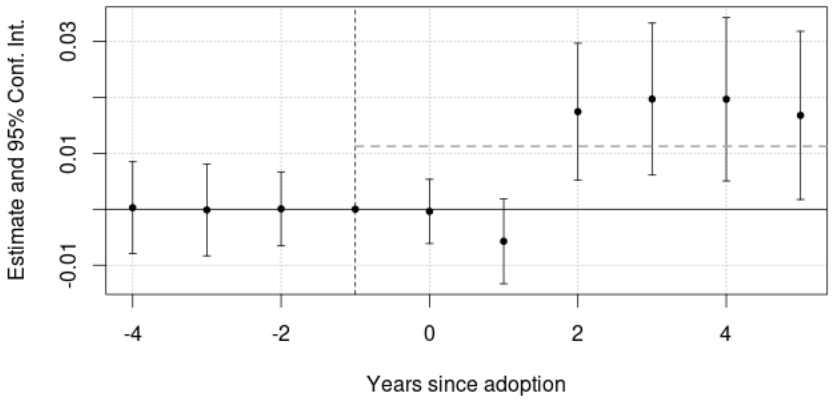
Note: Points report event-time estimates from the stacked difference-in-differences specification, and vertical bars show 95% confidence intervals. Event time is measured relative to the adoption year. The vertical dashed line marks the omitted reference period (k= -1). The solid horizontal line at zero indicates no estimated difference relative to that reference period.

Figure 5. Stacked DiD Estimates of Algorithm Adoption's Effect on Substantiation Rates for Low-Yield (Informal) Referrals



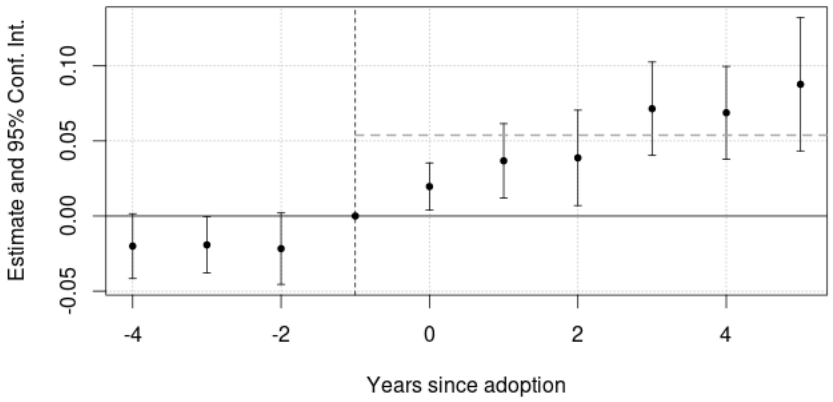
Note: Points report event-time estimates from the stacked difference-in-differences specification, and vertical bars show 95% confidence intervals. Event time is measured relative to the adoption year. The vertical dashed line marks the omitted reference period (k= -1). The solid horizontal line at zero indicates no estimated difference relative to that reference period.

Figure 6. Stacked DiD Estimates of Algorithm Adoption's Effect on Re-Referral Rates



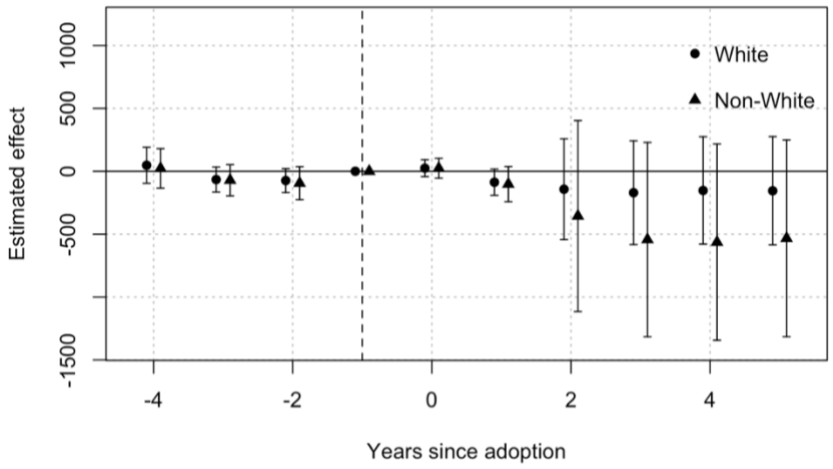
Note: Points report event-time estimates from the stacked difference-in-differences specification, and vertical bars show 95% confidence intervals. Event time is measured relative to the adoption year. The vertical dashed line marks the omitted reference period (k= -1). The solid horizontal line at zero indicates no estimated difference relative to that reference period. The gray horizontal dashed line denotes the average post-adoption effect across event years k=0 to k=5.

Figure 7. Stacked DiD Estimates of Algorithm Adoption's Effect on Risk-Score Gradient in Investigation Decisions



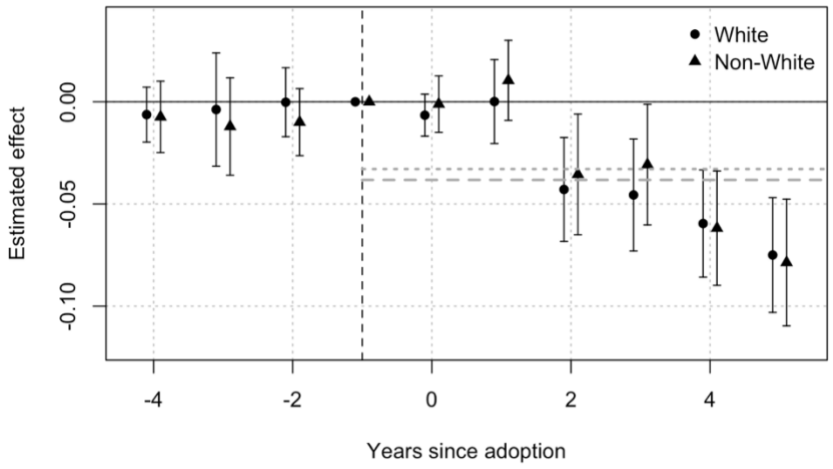
Note: Points report event-time estimates from the stacked difference-in-differences specification, and vertical bars show 95% confidence intervals. Event time is measured relative to the adoption year. The vertical dashed line marks the omitted reference period (k= -1). The solid horizontal line at zero indicates no estimated difference relative to that reference period. The gray horizontal dashed line denotes the average post-adoption effect across event years k=0 to k=5.

Figure 8. Stacked DiD Estimates of Algorithm Adoption's Effect on Investigation Counts by Race



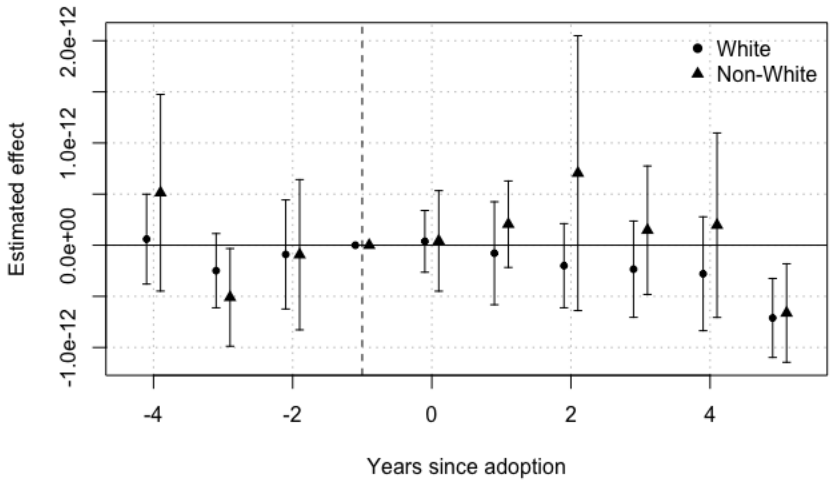
Note: Points report event-time estimates from the stacked difference-in-differences specification, and vertical bars show 95% confidence intervals. Event time is measured relative to the adoption year. The vertical dashed line marks the omitted reference period (k= -1). The solid horizontal line at zero indicates no estimated difference relative to that reference period. Circles denote estimates for white children, and triangles denote estimates for non-white children.

Figure 9. Stacked DiD Estimates of Algorithm Adoption's Effect on Substantiation Rates by Race



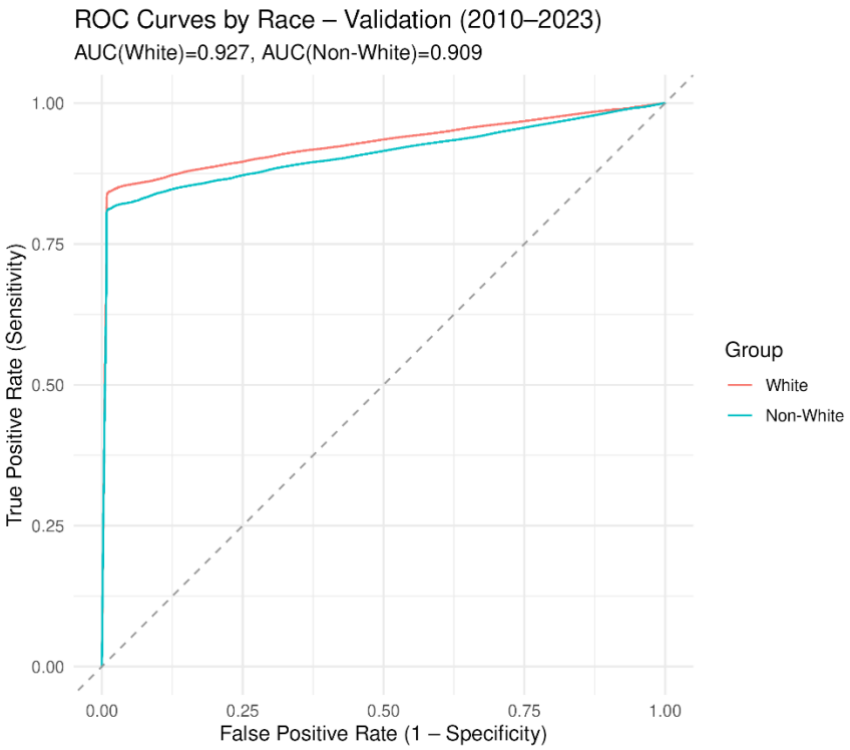
Note: Points report event-time estimates from the stacked difference-in-differences specification, and vertical bars show 95% confidence intervals. Event time is measured relative to the adoption year. The vertical dashed line marks the omitted reference period (k= -1). The solid horizontal line at zero indicates no estimated difference relative to that reference period. The gray horizontal dashed lines denote the group-specific average post-adoption effects across event years k = 0 to k = 5; the long-dashed line corresponds to white children, and the short-dashed line corresponds to non-white children. Circles denote estimates for white children, and triangles denote estimates for non-white children.

Figure 10. Stacked DiD Estimates of Racial Differences in Algorithmic Risk-Score Variation



Note: Points report event-time estimates from the stacked difference-in-differences specification, and vertical bars show 95% confidence intervals. Event time is measured relative to the adoption year. The vertical dashed line marks the omitted reference period (k= -1). The solid horizontal line at zero indicates no estimated difference relative to that reference period. Circles denote estimates for white children, and triangles denote estimates for non-white children.

Figure 11. ROC Curves by Race – Validation Sample (2010–2023)



Note: ROC curves plot the true positive rate against the false positive rate for the NCANDS-based pseudo-model, separately for white and non-white children. AUC values summarize predictive discrimination, with higher values indicating better ranking performance. The gray dashed diagonal represents random classification.

References

Administration for Children and Families. (2022). *Child Maltreatment 2022*. https://www.acf.hhs.gov/sites/default/files/documents/cb/cm2022.pdf

Administration for Children and Families. (2023). *New Child Maltreatment Report Finds Child Abuse and Neglect Decreased to a Five-Year Low*. https://www.acf.hhs.gov/media/press/2023/new-child-maltreatment-report-finds-child-abuse-and-neglect-decreased-five-year

Allegheny County. (2019). *Allegheny Family Screening Tool: Methodology*. https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/Methodology-V2-from-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-7.pdf

Alon-Barkat, S., & Busuioc, M. (2023). Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice. *Journal of Public Administration Research and Theory*, *33*(1), 153–169. https://doi.org/10.1093/jopart/muac007

Alon-Barkat, S., Busuioc, M., Schwoerer, K., & Weißmüller, K. S. (2025). Algorithmic discrimination in public service provision: Understanding citizens' attribution of responsibility for human versus algorithmic discriminatory outcomes. *Journal of Public Administration Research and Theory*, *35*(4), 469–488. https://doi.org/10.1093/jopart/muaf024

Angelova, V., Dobbie, W. S., & Yang, C. (2023). *Algorithmic recommendations and human discretion*. National Bureau of Economic Research. https://www.nber.org/papers/w31747

Arnold, D., Dobbie, W., & Hull, P. (2021). Measuring Racial Discrimination in Algorithms. *AEA Papers and Proceedings*, *111*, 49–54. https://doi.org/10.1257/pandp.20211080

Artinger, F. M., Artinger, S., & Gigerenzer, G. (2019). C. Y. A.: Frequency and causes of defensive decisions in public administration. *Business Research*, *12*(1), 9–25. https://doi.org/10.1007/s40685-018-0074-2

ASPE. (2017). *Predictive Analytics in Child Welfare*. ASPE. http://aspe.hhs.gov/predictive-analytics-child-welfare

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech Era. *Journal of Financial Economics*, *143*(1), 30–56. https://doi.org/10.1016/j.jfineco.2021.05.047

Basu, A. (2024). Mitigating Clinical Algorithmic Discrimination. *JAMA Health Forum*, *5*(12), e244190. https://doi.org/10.1001/jamahealthforum.2024.4190

Basu, A., Hammarlund, N., Khor, S., & Bansal, A. (2021). *Understanding Algorithmic Discrimination in Health Economics Through the Lens of Measurement Errors* (SSRN Scholarly Paper No. 3953940). https://papers.ssrn.com/abstract=3953940

Bello-Gomez, R. A. (2020). Interacting Capacities: The Indirect National Contribution to Subnational Service Provision. *Public Administration Review*, *80*(6), 1011–1023. https://doi.org/10.1111/puar.13192

Bertelli, A. M., & Jr, L. E. L. (2006). *Madison's Managers: Public Administration and the Constitution*. Johns Hopkins University Press.

Bouchard, G., & Carroll, B. W. (2002). Policy-making and administrative discretion: The case of immigration in Canada. *Canadian Public Administration*, *45*(2), 239–257. https://doi.org/10.1111/j.1754-7121.2002.tb01082.x

Bovens, M., & Zouridis, S. (2002). From Street-Level to System-Level Bureaucracies: How Information and Communication Technology is Transforming Administrative Discretion and Constitutional Control. *Public Administration Review*, *62*(2), 174–184. https://doi.org/10.1111/0033-3352.00168

Brayne, S., & Christin, A. (2021). Technologies of Crime Prediction: The Reception of Algorithms in Policing and Criminal Courts. *Social Problems*, *68*(3), 608–624. https://doi.org/10.1093/socpro/spaa004

Budzyń, K., Romańczyk, M., Kitala, D., Kołodziej, P., Bugajski, M., Adami, H. O., Blom, J., Buszkiewicz, M., Halvorsen, N., Hassan, C., Romańczyk, T., Holme, Ø., Jarus, K., Fielding, S., Kunar, M., Pellise, M., Pilonis, N., Kamiński, M. F., Kalager, M., … Mori, Y. (2025). Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: A multicentre, observational study. *The Lancet Gastroenterology & Hepatology*, *10*(10), 896–903. https://doi.org/10.1016/S2468-1253(25)00133-5

Buffat, A. (2015). Street-Level Bureaucracy and E-Government. *Public Management Review*, *17*(1), 149–161. https://doi.org/10.1080/14719037.2013.771699

Bullock, J. B. (2019). Artificial Intelligence, Discretion, and Bureaucracy. *The American Review of Public Administration*, *49*(7), 751–761. https://doi.org/10.1177/0275074019856123

Bullock, J., Young, M. M., & Wang, Y.-F. (2020). Artificial intelligence, bureaucratic form, and discretion in public service. *Information Polity*, *25*(4), 491–506. https://doi.org/10.3233/IP-200223

Busuioc, M. (2021). Accountable Artificial Intelligence: Holding Algorithms to Account. *Public Administration Review*, *81*(5), 825–836. https://doi.org/10.1111/puar.13293

Cheng, H.-F., Stapleton, L., Kawakami, A., Sivaraman, V., Cheng, Y., Qing, D., Perer, A., Holstein, K., Wu, Z. S., & Zhu, H. (2022). How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions. *CHI Conference on Human Factors in Computing Systems*, 1–22. https://doi.org/10.1145/3491102.3501831

Cheng, L., & Chouldechova, A. (2022). *Heterogeneity in Algorithm-Assisted Decision-Making: A Case Study in Child Abuse Hotline Screening* (arXiv:2204.05478). arXiv. https://doi.org/10.48550/arXiv.2204.05478

Chenot, D. (2011). The Vicious Cycle: Recurrent Interactions Among the Media, Politicians, the Public, and Child Welfare Services Organizations. *Journal of Public Child Welfare*, *5*(2–3), 167–184. https://doi.org/10.1080/15548732.2011.566752

Christensen, R. K., & Gazley, B. (2008). Capacity for public administration: Analysis of meaning and measurement. *Public Administration and Development*, *28*(4), 265–279. https://doi.org/10.1002/pad.500

Christin, A. (2017). Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society*, *4*(2), 2053951717718855. https://doi.org/10.1177/2053951717718855

Cohen, M. D., & Bacdayan, P. (1994). Organizational Routines Are Stored As Procedural Memory: Evidence from a Laboratory Study. *Organization Science*, *5*(4), 554–568.

Congressional Research Services. (2025). *Regulating Artificial Intelligence: U.S. and International Approaches and Considerations for Congress* [Legislation]. https://www.congress.gov/crs-product/R48555

Conrad, D., & Kellar-Guenther, Y. (2006). Compassion fatigue, burnout, and compassion satisfaction among Colorado child protection workers. *Child Abuse & Neglect*, *30*(10), 1071–1080. https://doi.org/10.1016/j.chiabu.2006.03.009

De-Arteaga, M., Fogliato, R., & Chouldechova, A. (2020). A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3313831.3376638

Drolc, C. A., & Keiser, L. R. (2021). The Importance of Oversight and Agency Capacity in Enhancing Performance in Public Service Delivery. *Journal of Public Administration Research and Theory*, *31*(4), 773–789. https://doi.org/10.1093/jopart/muaa055

Elston, T., & Zhang, Y. (2025). Ready, willing, and able? Bureaucratic capacity, slack resources, and political control. *Journal of Public Administration Research and Theory*, *35*(4), 452–468. https://doi.org/10.1093/jopart/muaf021

Espeland, W. N., & Sauder, M. (2007). Rankings and Reactivity: How Public Measures Recreate Social Worlds. *American Journal of Sociology*, *113*(1), 1–40. https://doi.org/10.1086/517897

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

Feldman, M. S., & Pentland, B. T. (2003). Reconceptualizing Organizational Routines as a Source of Flexibility and Change. *Administrative Science Quarterly*, *48*(1), 94–118. https://doi.org/10.2307/3556620

Fenger, M., & Simonse, R. (2024). The implosion of the Dutch surveillance welfare state. *Social Policy & Administration*, *58*(2), 264–276. https://doi.org/10.1111/spol.12998

Fong, K. (2020). Getting Eyes in the Home: Child Protective Services Investigations and State Surveillance of Family Life. *American Sociological Review*, *85*(4), 610–638. https://doi.org/10.1177/0003122420938460

Font, S., & Maguire-Jack, K. (2021). The Organizational Context of Substantiation in Child Protective Services Cases. *Journal of Interpersonal Violence*, *36*(15–16), 7414–7435. https://doi.org/10.1177/0886260519834996

Fryer Jr, G. E., Miyoshi, T. J., & Thomas, P. J. (1989). The relationship of child protection worker attitudes to attrition from the field. *Child Abuse & Neglect*, *13*(3), 345–350.

Gillingham, C., Morley, J., & Floridi, L. (2024). *The Effects of AI on Street-Level Bureaucracy: A Scoping Review* (SSRN Scholarly Paper No. 4823175). https://doi.org/10.2139/ssrn.4823175

Gillingham, P. (2019). Can Predictive Algorithms Assist Decision-Making in Social Work with Children and Families? *Child Abuse Review*, *28*(2), 114–126. https://doi.org/10.1002/car.2547

Gillingham, P., & Humphreys, C. (2010). Child Protection Practitioners and Decision-Making Tools: Observations and Reflections from the Front Line. *The British Journal of Social Work*, *40*(8), 2598–2616. https://doi.org/10.1093/bjsw/bcp155

Grimmelikhuijsen, S., & Meijer, A. (2022). Legitimacy of Algorithmic Decision-Making: Six Threats and the Need for a Calibrated Institutional Response. *Perspectives on Public Management and Governance*, *5*(3), 232–242. https://doi.org/10.1093/ppmgov/gvac008

Hauser, O. P., Light, M., Shelmerdine, L., & Blumenau, J. (2025). Why evaluating the impact of AI needs to start now. *Nature*, *643*(8073), 910–912. https://doi.org/10.1038/d41586-025-02266-7

Hood, C. (2011). *The Blame Game: Spin, Bureaucracy, and Self-Preservation in Government*. Princeton University Press.

Ingraham, P. W., Joyce, P. G., & Donahue, A. K. (2003). *Government Performance: Why Management Matters*. Johns Hopkins University Press.

Inman, C., & Adams, A. (2026). *Modernizing Child Welfare Technologies and Tools: Opportunities for Predictive Risk Modeling to Improve Child Safety and Outcomes*. https://acf.gov/sites/default/files/documents/main/Modernizing-Child-Welfare-Technologies-and-Tools-2026.3.5_508.pdf

Jagannathan, R., & Camasso, M. J. (2011). The crucial role played by social outrage in efforts to reform child protective services. *Children and Youth Services Review*, *33*(6), 894–900. https://doi.org/10.1016/j.childyouth.2010.12.010

Jagannathan, R., & Camasso, M. J. (2017). Social outrage and organizational behavior: A national study of child protective service decisions. *Children and Youth Services Review*, *77*, 153–163. https://doi.org/10.1016/j.childyouth.2017.03.015

Johnson, M. S., Levine, D. I., & Toffel, M. W. (2023). Improving Regulatory Effectiveness through Better Targeting: Evidence from OSHA. *American Economic Journal: Applied Economics*, *15*(4), 30–67. https://doi.org/10.1257/app.20200659

Kawakami, A., Sivaraman, V., Cheng, H.-F., Stapleton, L., Cheng, Y., Qing, D., Perer, A., Wu, Z. S., Zhu, H., & Holstein, K. (2022). *Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support* (arXiv:2204.02310). arXiv. https://doi.org/10.48550/arXiv.2204.02310

Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at Work: The New Contested Terrain of Control. *Academy of Management Annals*, *14*(1), 366–410. https://doi.org/10.5465/annals.2018.0174

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human Decisions and Machine Predictions*. *The Quarterly Journal of Economics*, *133*(1), 237–293. https://doi.org/10.1093/qje/qjx032

Kleinberg, J., & Raghavan, M. (2021). Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, *118*(22), e2018340118. https://doi.org/10.1073/pnas.2018340118

Leake, R., Rienks, S., & Obermann, A. (2017). A Deeper Look at Burnout in the Child Welfare Workforce. *Human Service Organizations: Management, Leadership & Governance*, 1–11. https://doi.org/10.1080/23303131.2017.1340385

LeBlanc, V. R., Regehr, C., Shlonsky, A., & Bogo, M. (2012). Stress responses and decision making in child protection workers faced with high conflict situations. *Child Abuse & Neglect*, *36*(5), 404–412. https://doi.org/10.1016/j.chiabu.2012.01.003

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*(2), 255–275. https://doi.org/10.1037/0033-2909.125.2.255

Levy, K., Chasalow, K. E., & Riley, S. (2021). Algorithms and Decision-Making in the Public Sector. *Annual Review of Law and Social Science*, *17*(Volume 17, 2021), 309–334. https://doi.org/10.1146/annurev-lawsocsci-041221-023808

Lipsky, M. (1980). *Street Level Bureaucracy: Dilemmas of the Individual in Public Services*. Russell Sage Foundation. https://www.jstor.org/stable/10.7758/9781610447713

Ludwig, J., Mullainathan, S., & Rambachan, A. (2024). *The Unreasonable Effectiveness of Algorithms*. https://www.nber.org/system/files/working_papers/w32125/w32125.pdf

Maguire-Jack, K., Font, S. A., & Dillard, R. (2020). Child Protective Services Decision-Making: The Role of Children's Race and County Factors. *The American Journal of Orthopsychiatry*, *90*(1), 48–62. https://doi.org/10.1037/ort0000388

March, J. G., & Simon, H. A. (1958). *Organizations*. Wiley-Blackwell.

Marshall, D. B., & English, D. J. (2000). Neural network modeling of risk assessment in child protective services: Psychological Methods. *Psychological Methods*, *5*(1), 102–124. (2000-03131-006). https://doi.org/10.1037/1082-989X.5.1.102

McFadden, P., Mallett, J., & Leiter, M. (2018). Extending the two-process model of burnout in child protection workers: The role of resilience in mediating burnout via organizational factors of control, values, fairness, reward, workload, and community relationships. *Stress and Health*, *34*(1), 72–83. https://doi.org/10.1002/smi.2763

Meijer, A., & Wessels, M. (2019). Predictive Policing: Review of Benefits and Drawbacks. *International Journal of Public Administration*, *42*(12), 1031–1039. https://doi.org/10.1080/01900692.2019.1575664

Mergel, I., Dickinson, H., Stenvall, J., & Gasco, M. (2023). Implementing AI in the public sector. *Public Management Review*, *0*(0), 1–14. https://doi.org/10.1080/14719037.2023.2231950

Merritt, D. H. (2020). How Do Families Experience and Interact with CPS? *The ANNALS of the American Academy of Political and Social Science*, *692*(1), 203–226. https://doi.org/10.1177/0002716220979520

Mohler, G. O., Short, M. B., Malinowski, S., Johnson, M., Tita, G. E., Bertozzi, A. L., & Brantingham, P. J. (2015). Randomized Controlled Field Trials of Predictive Policing. *Journal of the American Statistical Association*, *110*(512), 1399–1411. https://doi.org/10.1080/01621459.2015.1077710

Moynihan, D., Hybschmann, M., Gimborys, K., Loudin, S., & McClellan, W. (2025). Digital administrative burdens and clawback logics: The Australian Robodebt scheme. *Perspectives on Public Management and Governance*, gvaf011. https://doi.org/10.1093/ppmgov/gvaf011

Nelson, R. R., & Winter, S. G. (1982). *An evolutionary theory of economic change*. harvard university press. https://books.google.com/books?hl=en&lr=&id=6Kx7s_HXxrkC&oi=fnd&pg=PA1&dq=An+Evolutionary+Theory+of+Economic+Change&ots=7y5SODD4IJ&sig=CDwMF87ERIuLDCnd0O5MDV_tyjs

Nielsen, P. A., & Moynihan, D. P. (2017). How Do Politicians Attribute Bureaucratic Responsibility for Performance? Negativity Bias and Interest Group Advocacy. *Journal of Public Administration Research and Theory*, *27*(2), 269–283. https://doi.org/10.1093/jopart/muw060

Nohria, N., & Gulati, R. (1997). What is the optimum amount of organizational slack?: A study of the relationship between slack and innovation in multinational firms. *European Management Journal*, *15*(6), 603–611. https://doi.org/10.1016/S0263-2373(97)00044-3

OASH. (2024). *National Child Abuse and Neglect Data System (NCANDS)—Healthy People 2030 | health.gov*. https://health.gov/healthypeople/objectives-and-data/data-sources-and-methods/data-sources/national-child-abuse-and-neglect-data-system-ncands

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Parasuraman, R., & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors*, *52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Porter, T. M. (1995). *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton University Press. https://www.jstor.org/stable/j.ctt7sp8x

Rahman, H. A. (2021). The Invisible Cage: Workers' Reactivity to Opaque Algorithmic Evaluations. *Administrative Science Quarterly*, *66*(4), 945–988. https://doi.org/10.1177/00018392211010118

Regehr, C., Bogo, M., Shlonsky, A., & LeBlanc, V. (2010). Confidence and professional judgment in assessing children's risk of abuse. *Research on Social Work Practice*, *20*(6), 621–628. https://doi.org/10.1177/1049731510368050

Rinta-Kahila, T., Penttinen, E., Salovaara, A., Soliman, W., & Ruissalo, J. (2023). The Vicious Circles of Skill Erosion: A Case Study of Cognitive Automation. *Journal of the Association for Information Systems*, *24*(5), 1378–1412. https://doi.org/10.17705/1jais.00829

Rittenhouse, K., Putnam-Hornstein, E., & Vaithianathan, R. (2022). *ALGORITHMS, HUMANS AND RACIAL DISPARITIES IN CHILD PROTECTIVE SERVICES: EVIDENCE FROM THE ALLEGHENY FAMILY SCREENING TOOL*. https://krittenh.github.io/katherine-rittenhouse.com/Rittenhouse_Algorithms.pdf

Samant, A., Horowitz, A., Xu, K., & Beiers, S. (2021). *Family Surveillance by Algorithm*. https://www.aclu.org/documents/family-surveillance-algorithm

Schwartz, I. M., York, P., Nowakowski-Sims, E., & Ramos-Hernandez, A. (2017). Predictive and prescriptive analytics, machine learning and child welfare risk assessment: The Broward County experience. *Children and Youth Services Review*, *81*, 309–320. https://doi.org/10.1016/j.childyouth.2017.08.020

Sele, D., & Chugunova, M. (2024). Putting a human in the loop: Increasing uptake, but decreasing accuracy of automated decision-making. *PLOS ONE*, *19*(2), e0298037. https://doi.org/10.1371/journal.pone.0298037

Selten, F., Robeer, M., & Grimmelikhuijsen, S. (2023). 'Just like I thought': Street-level bureaucrats trust AI recommendations if they confirm their professional judgment. *Public Administration Review*, *83*(2), 263–278. https://doi.org/10.1111/puar.13602

Simon, H. A. (1947). *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization*. Free Pr.

Snow, T. (2021). From satisficing to artificing: The evolution of administrative decision-making in the age of the algorithm. *Data & Policy*, *3*, e3. https://doi.org/10.1017/dap.2020.25

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science (New York, N.Y.)*, *333*(6043), 776–778. https://doi.org/10.1126/science.1207745

Sprang, G., Craig, C., & Clark, J. (2011). Secondary traumatic stress and burnout in child welfare workers. *Child Welfare*, *90*(6), 149–168.

Stapleton, L., Cheng, H.-F., Kawakami, A., Sivaraman, V., Cheng, Y., Qing, D., Perer, A., Holstein, K., Wu, Z. S., & Zhu, H. (2022, April 29). *Extended Analysis of "How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions."* arXiv.Org. https://arxiv.org/abs/2204.13872v1

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, *8*(12), 2293–2303. https://doi.org/10.1038/s41562-024-02024-1

Vaithianathan, R., Kulick, E., Putnam-Hornstein, E., & Benavides-Prado, D. (2019). *Allegheny Family Screening Tool: Methodology, Version 2—Centre for Social Data Analytics—AUT*. Allegheny Family Screening Tool: Methodology, Version 2. https://csda.aut.ac.nz/research/our-publications/2019/allegheny-family-screening-tool-methodology,-version-2

Waardenburg, L., Huysman, M., & Sergeeva, A. V. (2022). In the Land of the Blind, the One-Eyed Man Is King: Knowledge Brokerage in the Age of Learning Algorithms. *Organization Science*, *33*(1), 59–82. https://doi.org/10.1287/orsc.2021.1544

Weaver, R. K. (1986). The Politics of Blame Avoidance. *Journal of Public Policy*, *6*(4), 371–398.

Wildeman, C., & Emanuel, N. (2014). Cumulative Risks of Foster Care Placement by Age 18 for U.S. Children, 2000–2011. *PLOS ONE*, *9*(3), e92785. https://doi.org/10.1371/journal.pone.0092785

Williams, M. J. (2021). Beyond state capacity: Bureaucratic performance, policy implementation and reform. *Journal of Institutional Economics*, *17*(2), 339–357. https://doi.org/10.1017/S1744137420000478

Wing, C., Freedman, S. M., & Hollingsworth, A. (2024). *Stacked Difference-in-Differences* (Working Paper No. 32054). National Bureau of Economic Research. https://doi.org/10.3386/w32054

Wu, X., Ramesh, M., & Howlett, M. (2015). Policy capacity: A conceptual framework for understanding policy competences and capabilities. *Policy and Society*, *34*(3–4), 165–171. https://doi.org/10.1016/j.polsoc.2015.09.001

Appendix 1. Robustness Checks on Selection and Resource Allocation

To test two key alternative explanations for our main findings, this section examines whether algorithm adoption coincided with changes in front-end case selection or agency resources. As the NCANDS Child File lacks direct measures of screened-out referrals or staffing, I construct a state-year panel from the NCANDS Agency File and replicate the main stacked difference-in-differences design at the state level[7].

First, to address the concern of selection bias in case selection. Figure A1 reports the event-time coefficients for the number of screened-out referrals. Pre-adoption estimates are jointly indistinguishable from zero, and post-adoption effects are also not significant. A marginal significant estimate at t=1 does not persist in adjacent periods and is not indicative of a systematic shift in intake selection. Second, to test for confounding resource shifts on the staff capacity after algorithm adoption. Figure A2 presents the corresponding event-study for screening, intake, and investigation staffing. Coefficients are centered near zero throughout the post-adoption window, indicating no detectable change in staffing capacity at the time of adoption.

These null effects provide evidence against alternative explanations rooted in either front-end selection bias or resource reallocation. Consequently, the main findings are more plausibly interpreted as reflecting changes in operational and analytical capacity rather than compositional shifts in cases or staffing.

[7] Because adoption in Pennsylvania occurred only in a single county, it is excluded from these state-level analyses.

Figure 1A. Stacked DiD Estimates of Algorithm Adoption's Effect on Screened-out Referrals

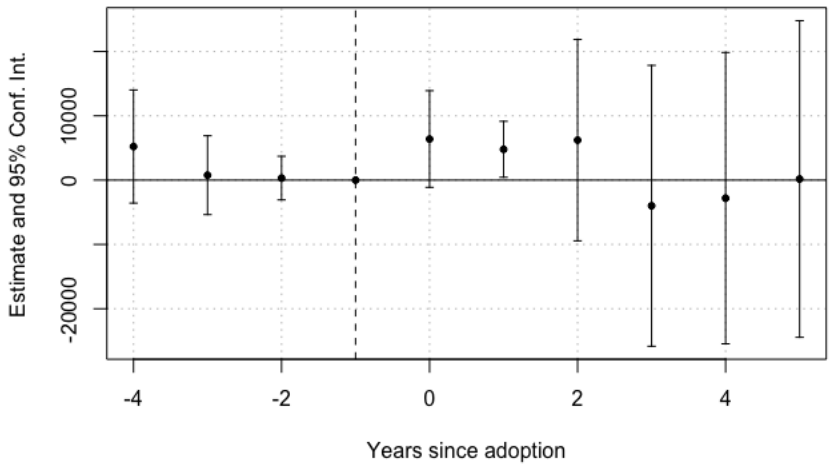
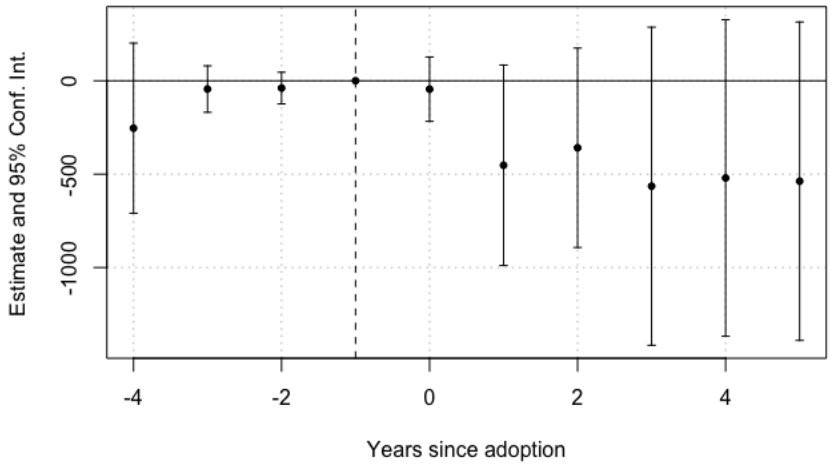


Figure 1B. Stacked DiD Estimates of Algorithm Adoption's Effect on Screening/Intake/Investigation Staffing



Appendix 2. Triple-DiD Estimates of Algorithm Adoption's Effect on Substantiation Rates by Race

	Triple DiD
event_time = -4 × treat	-0.021
	(0.008)
event_time = -3 × treat	-0.022
	(0.008)
event_time = -2 × treat	-0.016
	(0.006)
event_time = 0 × treat	-0.003
	(0.006)
event_time = 1 × treat	-0.003
	(0.006)
event_time = 2 × treat	0.005
	(0.009)
event_time = 3 × treat	0.017
	(0.008)
event_time = 4 × treat	-0.005
	(0.007)
event_time = 5 × treat	-0.005
	(0.009)
raceNonwhite	0.008
	(0.001)
raceWhite × event_time = -4 × treat	-0.007

	Triple DiD
	(0.004)
raceWhite × event_time = -3 × treat	0.001
	(0.005)
raceWhite × event_time = -2 × treat	0.001
	(0.006)
raceWhite × event_time = 0 × treat	-0.006
	(0.004)
raceWhite × event_time = 1 × treat	-0.011
	(0.006)
raceWhite × event_time = 2 × treat	-0.008
	(0.007)
raceWhite × event_time = 3 × treat	-0.016
	(0.006)
raceWhite × event_time = 4 × treat	-0.001
	(0.005)
raceWhite × event_time = 5 × treat	-0.002
	(0.005)
Num.Obs.	47804
R2 Within	0.015
R2 Within Adj.	0.014
Std.Errors	by: county
FE: county	Yes
FE: event time	Yes

Appendix 3. Lasso Regression Training Procedure and Validation Metrics for Replicated Screening Model

I estimated the screening model using the raw NCANDS Child File alone, covering reports filed from 2000 through 2009. After ordering each child's referrals by report date and retaining only the first record per child, the analytic file contained 18677739 unique investigations. I then split the data chronologically: reports from 2000–2007 (14609245 observations) formed the training set, while those from 2008–2009 (4068494 observations) constituted a strict hold-out validation set, yielding zero overlap in child identifiers across periods.

With the exception of child demographic characteristics such as gender and race, I included virtually every variable available in NCANDS. The predictors therefore span child risk factors (e.g., intellectual disability, emotional disturbance), caregiver risk factors (e.g., alcohol or drug abuse), and detailed perpetrator information. All categorical predictors were converted to factors with an explicit "missing" level; numeric variables received companion "_missing" flags and were median-imputed before one-hot encoding, after which the resulting sparse matrix was passed to a penalized logistic-regression estimator.

To select an optimal penalized parameter, I carried out an explicit grid search. A 3 percent stratified subsample of the training data served as a tuning pool. Within this subset, I fit five-fold cross-validated elastic-net models across five mixing parameters ($\alpha = 0, 0.25, 0.50, 0.75, 1$) and forty logarithmically spaced values of λ. The highest mean out-of-fold AUC (0.937) was achieved by a pure lasso ($\alpha = 1$) with $\lambda = 8.68 \times 10^{-5}$. Fixing those hyper-parameters, I re-estimated the model on the full training set and applied it to the hold-out cohort, where it attained an AUC of 0.931. This systematic search across multiple (α, λ) combinations ensures that the reported performance reflects the best-performing specification rather than an arbitrary choice.

Figure A displays the resulting ROC curve. Its steep ascent from the origin underscores the model's strong discriminative power in distinguishing substantiated from unsubstantiated investigations.

Figure 3A. ROC curve for the 2008–2009 hold-out sample (AUC = 0.931)