sociological science

How Do (Human) Child Welfare Workers Respond to Machine-Generated Risk Scores?

Martin Eiermann,[a] Maria Fitzpatrick,[b,c] Katharine Sadowski,[d] Christopher Wildeman[e,f,g]

a) University of Wisconsin-Madison; b) Cornell University; c) National Bureau of Economic Research; d) Stanford University; e) Duke University; f) Sanford School of Public Policy; g) ROCKWOOL Foundation Research Unit

Abstract: Algorithmic risk scoring tools have been widely incorporated into governmental decision making, yet little is known about how human decision makers interact with machine-generated risk scores at the street level. We examined such human–machine interactions in the child welfare system, a high-stakes setting where caseworkers ascertain whether government interventions in family life are warranted. Using novel data—verbatim transcripts of caseworker discussions—we found that decision makers: (1) disregarded scores in the middle of the distribution while paying attention to extremely high or low risk scores and (2) rationalized divergences between human decisions and machine-generated scores by highlighting the algorithm's overemphasis on historical data and specific risk factors and its lack of contextual knowledge. This meant that caseworkers were unlikely to modify their decisions so that they aligned with risk scores. However, we did not find evidence of principled resistance to algorithmic tools. Our findings advance research on such tools by specifying how human perceptions of the utility and limitations of novel technologies shape discretionary decision making by state officials; and they help to explain their uneven and potentially modest impact on the bureaucratic management of social vulnerability.

Keywords: child welfare; predictive algorithms; risk scoring algorithmic decision making; public administration; human–machine interactions

Reproducibility Package: The terms of our Data Use Agreement with the Douglas County Department of Human Services (DCDHS) legally prohibit us from sharing the original data, which are temporarily stored on a secure Cornell University research server, cannot be shared externally, and must be destroyed at the end of the agreement period. These restrictions reflect the presence of highly sensitive child welfare data in verbatim transcripts of caseworker discussions. All analysis code and documentation of qualitative coding workflows are publicly available at OSF. Researchers with questions about Douglas County Decision Aide (DCDA) data that were generated during the randomized controlled trial may contact: Ruby Richards, Director of Human Services, Douglas County (303-688-4825).

Citation: Eiermann, Martin, Maria Fitzpatrick, Katharine Sadowski, and Christopher Wildeman. 2025. "How Do (Human) Child Welfare Workers Respond to Machine-Generated Risk Scores?" Sociological Science 13: 1-21.

Received: September 3, 2025

Accepted: November 14, 2025

Published: January 6, 2026

Editor(s): Ari Adut, Jeremy Freese

DOI: 10.15195/v13.a1

EVERY day, individuals who work on the front lines of government agencies are charged with making difficult and consequential decisions. Police officers decide how to respond to 911 calls, judges set bail for defendants, parole boards determine if a person is released from prison, and child welfare caseworkers assess the need for state interventions in family life.

In recent years, those decisions have increasingly incorporated risk scores that are generated by so-called "predictive risk models" (PRM), which analyze statistical

patterns in historical administrative records and use these data to generate personalized numeric scores that forecast the risk of some future (and almost always unwanted) outcome, such as the likelihood of recidivism, hospital readmission, or child maltreatment (Brayne 2017; Burrell and Fourcade 2021; Green and Chen 2021; Kleinberg et al. 2018a; Neil and Zanger-Tishler 2025; Saeed, Patel, and Odeyemi 2022). Such novel tools are commonly integrated into administrative decision making through an approach known as "algorithm-in-the-loop": They do not simply replace human judgment but augment it (Brayne and Christin 2021; Burton, Stein, and Jensen 2020; Green and Chen 2019a, 2019b; Mahmud et al. 2022; Pruss 2023; Stevenson and Doleac 2024). Put in slightly different terms, machine-generated risk scores routinely intersect with the reasoning of human decision makers (such as judges or child welfare caseworkers) who have deep experience and deep-seated priors, who operate within organizationally bounded decision-making frameworks, and whose discretionary decisions can substantively impact program implementation and client outcomes (Evans and Harris 2004; Lipsky 2010; Maynard-Moody and Musheno 2012).

Several studies have tested how the adoption of PRM tools shapes decision-making patterns in public administration. They found that the algorithmic augmentation of administrative decision making can affect the overall incidence and the racial disparities of welfarist and penal interventions (Bhatt et al. 2024; Cheng et al. 2022; Eiermann 2024; Grimon and Mills 2025; Kleinberg et al. 2018a; Parker et al. 2022) but that impacts and efficiency gains can be uneven, unexpectedly modest, or statistically insignificant (Fitzpatrick, Sadowski, and Wildeman 2025; Imai et al. 2023; Stevenson 2018). This may be because the algorithm-in-the-loop approach still requires final decisions to be made by human officials, who do not necessarily update their priors when confronted with quantified metrics, regard new algorithmic tools as threats to their expertise and autonomy, or override algorithmic recommendations by over-indexing on highly salient characteristics (Agarwal et al. 2024; Brayne and Christin 2021; Green and Chen 2019b; Mullainathan and Obermeyer 2022; Skeem, Scurich, and Monahan 2020; Stevenson and Doleac 2024). Understanding how human–machine interactions unfold in different administrative agencies therefore holds particular promise for sociological research because it allows scholars to connect knowledge about the *impacts* of algorithmic tools on governmental decision making to insights into their *deployment* at the street level (Airoldi 2022; Stevenson and Doleac 2024).

However, those quotidian interactions are rarely observed, recorded, or otherwise documented in a systematic fashion. Although researchers are often able to collect comprehensive data on downstream outcomes such as criminal sentencing or parole decisions, child maltreatment investigations, or child hospitalizations—which are potentially influenced by PRM tools (Berk 2017; Bhatt et al. 2024; Duwe and Kim 2017; Eiermann 2024; Grimon and Mills 2025; Montana et al. 2023)—the upstream integration of human expertise and machine-generated metrics into a single decision-making loop has proven difficult to study. This is one reason why theoretical perspectives on algorithmic governance tend to have more to say about fairness and inequality at the population level than about human–machine interactions and discretionary decision making at the street level.

In this study, we offer a close-up view of human–machine interactions by focusing on a particularly consequential domain of public administration: the American child welfare system. Each year, more than seven million children come into contact with Child Protective Services (CPS) because they are reported for suspected abuse or neglect (U.S. Department of Health & Human Services, Administration for Children and Families, Administration on Children, Youth and Families, Children's Bureau 2024). CPS caseworkers must then make difficult decisions about state interventions in family life that are rendered even more complicated by incomplete or contradictory evidence, linguistic challenges when dealing with non-native speakers, the compounding hardships that welfare-involved families often face, and the inherent difficulties of ascertaining the true extent of past maltreatment and future harm for vulnerable children. Although broad policy frameworks and funding priorities are set at the state level, local CPS offices retain considerable discretion over the screening and investigation of maltreatment reports. This devolved decision making makes it particularly important to understand the frontline deployment of novel technologies, which are now used by CPS in many U.S. counties, including most major metropolitan areas (Eiermann 2024; Eubanks 2018; Saxena et al. 2020).

We proceed by leveraging a highly unique data source: verbatim transcripts of child welfare triage meetings. During these meetings, teams of caseworkers called RED teams (read, evaluate, and direct teams) reviewed incoming reports of alleged maltreatment and either recommended to "screen-out" a referral (i.e., no further action was taken) or "assign" it to a CPS caseworker for investigation. We were able to obtain data access from the Douglas County, Colorado Department of Human Services (DCDHS) as part of a larger randomized controlled trial (RCT), whose purpose was to evaluate a newly designed PRM tool called the Douglas County Decision Aide (DCDA). Our qualitative data collection began soon after the introduction of DCDA into the county's structured decision-making (SDM) process (Vaithianathan et al. 2019), with the explicit aim of understanding how frontline staff used the tool during a transitional period of organizational decision making. The data cover 207 separate maltreatment reports, of which a subset was randomly assigned to a RED team that had access to DCDA. The tool displayed two pieces of information: a risk score (scaled 1–20) that expressed the predicted risk level of a child's placement in out-of-home care within two years, computed from lasso regularized regressions, and the predicted percentages of children experiencing such out-of-home placement at each ventile (Figure 1).

Our findings show that human–machine interactions are shaped by the perceived utility and limitations of algorithmic tools, not by the availability of those tools per se. Specifically, we demonstrate that caseworkers paid particular attention to extremely high and low risk scores, which were seen as most informative, while discounting scores in the middle of the risk distribution as well as scores that diverged from their own assessments of maltreatment reports. We identified only two instances of caseworkers changing their decisions in response to high risk scores. Taken together, these findings also yield a potential explanation for the uneven and unexpectedly modest impact of algorithmic tools on administrative decision making observed in prior studies and the larger RCT (Fitzpatrick et al. 2025): When emerging technologies are embedded into established institutional processes, the discounting of divergent or uninformative scores imposes an upper bound on their

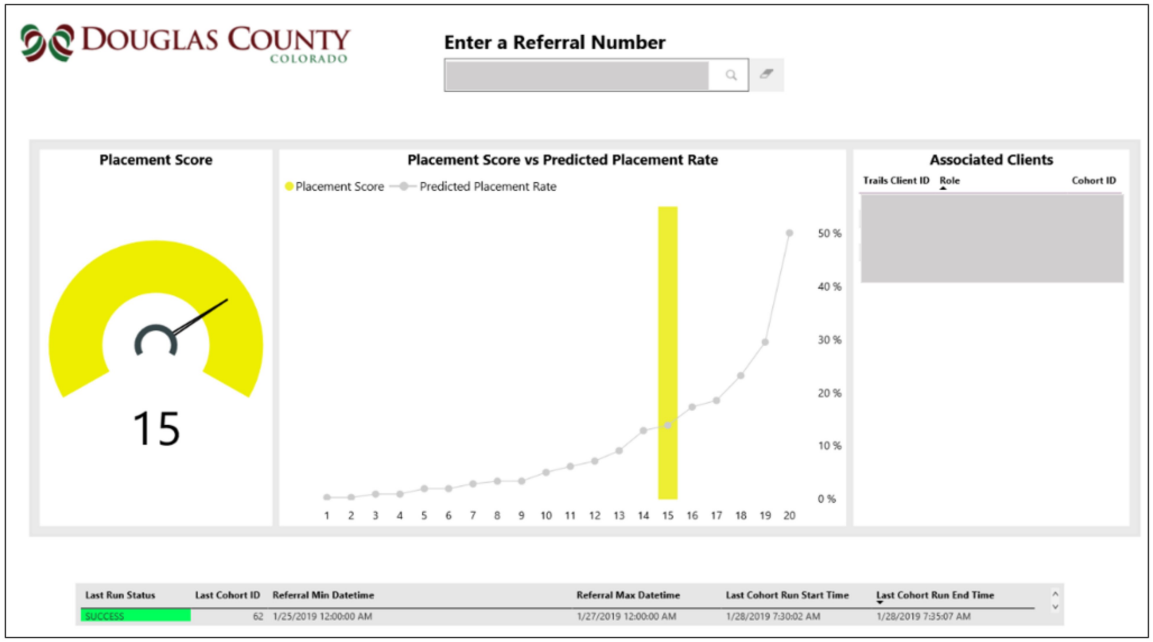

Figure 1: DCDA computer screenshot, as seen by RED team members. Information about the child and referral has been redacted. On the left-hand side of the screen is the score with a dial representing its place in the range of possible scores. In the middle of the screen is a figure showing the placement score and the related predicted placement rate at that score relative to predicted placement rates throughout the score distribution.

impact even when the use of algorithmic tools is mandated by CPS policy and when street-level administrators embrace the logic of ordinalization during their daily triage decisions (Fourcade and Healy 2024).

Administrative Decision Making in the Ordinal Society

The recent proliferation of algorithmic tools is the latest chapter in the quantification of administrative decision making (Desrosières 2002; Espeland and Stevens 2008; Koopman 2019; Porter 1995), substituting "technical efficiency and mechanical objectivity" for idiosyncratic forms of official judgment (Burrell and Fourcade 2021, P. 221). Machine-generated risk scores promise to facilitate "the comparison of different entities according to a common metric" (Espeland and Stevens 1998, P. 313) and to catalyze the Weberian transition from decisions "moved by personal sympathy and favor, by grace and gratitude" toward "calculable rules" that are applied "without regard for persons" and "with as much speed as possible" (Weber 1978, P. 974–975). But, more specifically, the proliferation of such scores also hints at the shifting nature of governance in the so-called "ordinal society" (Fourcade and Healy 2024)—a social formation organized around technologies that translate facts of everyday life into discrete, standardized units of information, use this information to populate hierarchically ordered rankings, and then leverage those rankings to tailor interventions toward differently scored individuals. The aim is not to produce *equal* treatments of those who come into contact with governing

agencies by intervening in all scenarios but *appropriately unequal* treatments based on historical and behavioral patterns and predicted future risk (Accominotti 2021; Amoore 2009; Cheney-Lippold 2011; Fourcade 2021; Fourcade and Gordon 2020; Mau 2019).

Child welfare systems exemplify this transformation, as many CPS agencies have adopted PRMs that use data on prior family contact with governing agencies to quantify the predicted likelihood of future maltreatment investigations, the future removal of a child from the home, or the likelihood of family reunification after a foster care placement, each of which can then inform intervention strategies and placement decisions (Cheng et al. 2022; Cuccaro-Alamin et al. 2017; Eiermann 2024; Eubanks 2018; Parker et al. 2022; Saxena et al. 2020; Vaithianathan et al. 2019). The appeal of such models is rooted in what Theodore Porter has called "mechanical objectivity" (Porter 1995): By eliminating personal bias and arbitrary judgment through seemingly objective measures of risk, officials hope to improve the allocation of scarce governmental resources while increasing the perceived legitimacy of CPS interventions in ostensibly "private" domains, such as family life (Brayne 2020; Burrell and Fourcade 2021; Kleinberg et al. 2018b).

This transformation has opened up the possibility of "cyberdelegation" (Cuellar 2016, 2017): the consignment of administrative agency from human caseworkers to algorithmic tools, exemplified most recently by the introduction of increasingly capable and autonomously operating artificial intelligence (AI) agents. In practice, however, algorithmic tools also remain embedded into larger socio-political systems and existing organizational structures (Coglianese and Dor 2020; Espeland and Stevens 2008). Street-level administrators engage with machine-generated scores while also drawing on their professional expertise and personal judgment, often with an eye toward navigating practical constraints within a system of formalized rules and semi-structured processes (Ellis, Davis, and Rummery 1999). This is widely seen as the price of "getting the job done" (Lipsky 2010, P. 18) while also creating opportunities for the "creative appropriation" of risk scores (Beraldo and Milan 2019, P. 3) or subtle forms of resistance against new technologies (Brayne and Christin 2021).

Because the use of algorithmic tools involves ongoing negotiation and adaptation rather than straightforward technical implementation (Espeland and Stevens 2008; Evans and Harris 2004; Seaver 2018), it is therefore particularly important to understand "how man and machine interact" (Stevenson and Doleac 2024, P. 382). Such interactions can potentially explain why the practical impact of PRM tools is uneven across jurisdictions and organizations, more modest than anticipated, limited to the short term, or correlated with socio-demographic indicators that are not included in algorithmic input data but are readily apparent to human decision makers, such as racial identity or socio-economic status (Eiermann 2024; Fitzpatrick et al. 2025; Imai et al. 2023; Skeem et al. 2020; Sloan, Naufal, and Caspers 2025; Stevenson 2018; Stevenson and Doleac 2024). Prior works have discussed several distinct forms of human–machine interaction as follows:

1. The introduction of novel algorithmic tools can displace idiosyncratic and local logics of administrative decision making when state officials defer to machine-generated scores (Espeland and Stevens 2008; Fourcade and Gordon

2020). Officials are more likely to *embrace* such scores when they are perceived as more accurate than human assessments, when they provide information that is considered useful but was previously inaccessible, and when they make it easier to manage administrative workloads (Kleinberg et al. 2018b).

2. Street-level administrators may accept the paradigm (and/or embrace the promise) of ordinalization but nonetheless consider PRMs as potentially flawed and experimental technologies (Cheng et al. 2022), leading to compensatory strategies that leverage local expertise to *adjust* for the perceived flaws of decontextualized machine-generated knowledge (Preston-Shoot 2001).

3. State officials may *ignore* risk scores that are not perceived as useful or accurate. When policy frameworks or administrative rules and regulations nonetheless mandate their use, officials must then perform acts of symbolic compliance even if they make no substantial changes to their internal practices and do not consider machine-generated scores as important elements of their decision making (Stevenson and Doleac 2024).

4. Officials may *resist* algorithmic tools if they threaten the autonomy or expertise of trained professionals, leading to a variety of strategies that can range from foot dragging (i.e., delaying the implementation of such scores) to overt attempts at discouraging the routine use of novel technologies (Brayne and Christin 2021; Merry 2016).

Knowledge of such human–machine interactions becomes particularly important when organizational structures are decentralized and workloads are high, as is the case in the U.S. child welfare system (Edwards and Wildeman 2018; Scott 1999). Yet relevant empirical research remains relatively sparse and is overwhelmingly focused on law enforcement agencies and courts (Brayne 2017, 2020; Brayne and Christin 2021; Green and Chen 2019b; Stevenson 2018; Stevenson and Doleac 2024). However, the "humans" who interact with "machines" in criminal justice settings (e.g., judges and police officers) differ substantially from most U.S. public servants: they are uniformed (or, in the case of judges, robed), sometimes armed, comparatively well-paid, and protected with considerable status privileges. Decision-making processes in criminal justice settings also differ. Law enforcement officers must make decisions faster and more spontaneously than child welfare caseworkers, whose workflows are often governed by formal SDM processes and potentially include one or more layers of supervisory review.

This scarcity of empirical data has complicated efforts to substantiate the existence (or evaluate the importance) of various forms of human–machine interactions beyond the criminal justice system and across the multitude of agencies that collectively constitute the "many-handed" American state (Morgan and Orloff 2017). This includes the child welfare system. With the exception of one study that used retrospective interviews with CPS caseworkers to understand how machine-generated scores were used to address persistent racial disparities in child welfare decision making (Cheng et al. 2022), prior research has largely focused on measuring the predictive fairness of CPS algorithms and assessing their impact on child welfare outcomes (Cuccaro-Alamin et al. 2017; Eiermann 2024; Eubanks 2018; Parker et al.

2022; Saxena et al. 2020). By identifying patterns of human–machine interactions in real-world CPS settings, we produce domain-specific knowledge about technological innovation in the child welfare system and also contribute more generally to sociological understandings of public administration in the ordinal society.

Data and Methods

Research Design and Data Collection

Our data come from CPS referral screenings in Douglas County, CO, where officials had introduced DCDA, a new risk scoring tool that can be used to augment caseworker screenings of incoming maltreatment reports. DCDA uses lasso regularized regressions to analyze 460 predictors and returns a numeric score between 1 and 20 that predicts a child's risk of being removed from the home within 24 months of a referral. At the time of data collection, children with scores less than 9 had a less than 5 percent chance of experiencing out-of-home placement within 24 months, but children with a score of 20 had a 35 percent chance of experiencing out-of-home placement.

The model was trained on administrative data from the Denver metro area—excluding Douglas County—using records from a Statewide Automated Child Welfare Information System and public welfare eligibility data (Vaithianathan et al. 2019). Unlike other risk scoring tools used by CPS (Eubanks 2018), DCDA did not factor in parental criminal-justice data. The DCDHS developed this tool in cooperation with the Centre for Social Data Analytics at the Auckland University of Technology. None of the authors of this study were involved in this development.

DCDHS randomly assigned RED teams to either access the DCDA score during their decision-making meeting or not access the score, and then agreed to videotape RED team meetings for this study. 52 meetings were videotaped, but only 41 of those recordings were properly saved and could be transcribed by trained research assistants. Each recording included information on 1–10 referrals. In total, our transcripts cover 207 referrals, or roughly five referrals per meeting on average. Each transcript was then hand coded by research assistants to identify the time spent on case reviews and deliberations, and coding discrepancies were adjudicated by senior members of the research team. We were able to consistently identify speakers within meetings based on ad hoc identifiers, such as clothing, but were unable to trace statements back to individual caseworkers across multiple meetings.

Research Methods

We analyzed these transcripts in three ways. First, we identified statistical patterns across all referrals, including the length of time spent on each referral. Second, we qualitatively analyzed transcripts, using an iterative open coding process that aggregated individual caseworker remarks into thematic clusters. We independently performed the qualitative analyses twice and used memos to resolve disagreements and identify core themes. Third, we tested the association between screening decisions and risk scores with a logit regression model.

Studies of human–machine interactions are complicated by the fact that PRM tools can evolve quickly and consequentially through updates to the underlying mathematical models and that changes to organizational processes can affect how frontline workers incorporate risk scores into their decision making. DCDA has been updated since its initial deployment, and RED teams shifted to virtual meetings during the COVID-19 pandemic. Overall screen-in rates in Douglas County also declined by 8–10 percent between 2020 and 2025 as CPS worked to address problems related to the over-reporting of maltreatment allegations in families of color and their subsequent over-representation in child welfare investigations. As a result, findings derived from this study may not reflect human–machine interactions with differently calibrated PRM tools and in other organizational settings.

The external validity of our findings also remains limited, because data come from one U.S. county only. However, working with transcript data confers two unique advantages: First, these data allowed us to study human–machine interactions directly, complementing quantitative research that analyzed the effects of risk scores on patterns of CPS interventions (Eiermann 2024; Fitzpatrick et al. 2025; Parker et al. 2022). Second, unlike data obtained from vignette-based or interview-based studies (Cheng et al. 2022; Green and Chen 2019a, 2019b, 2021; Lin et al. 2020), CPS transcript data capture actual deliberations rather than hypothetical scenarios or retrospective reflections.

It remains possible that caseworkers acted differently in recorded meetings than in non-recorded meetings, because they knew that their conversations would be analyzed by researchers. ("Oh, so this is part of the research?"/"Yes, this is why they're watching us.") This has the potential to bias our findings. However, caseworkers did not know what questions researchers would aim to answer with the recordings. We also considered selective self-censoring unlikely in light of caseworkers' small talk about a wide range of personal interests, often used to lighten the mood during incredibly challenging conversations about maltreatment referrals.

DCDA in Action

All RED team caseworkers received a mandatory 3 h training about DCDA and were advised that risk scores were computed from historical data on child welfare referrals and public benefits access, did not factor in information about the current referral, and did not mandate specific actions during the initial triage of maltreatment reports (Vaithianathan et al. 2019). Caseworkers then reviewed each referral using a well-established and consensus-based SDM process. First, they accessed historical data about parents' prior criminal justice contact and about children's prior CPS contact ("priors"). Second, they reviewed information from the maltreatment report ("report"), which had been provided by the person submitting the report to CPS—such as a teacher, school nurse, healthcare worker, police officer, or family member. Third, they deliberated as a group before deciding whether to suggest the report be screened out or assigned for investigation ("deliberation"). If a referral was discussed by a DCDA team as part of the RCT treatment group, caseworkers were instructed to access the risk score after concluding their initial discussion and resulting assessment, and then factor it into their final decision (Vaithianathan et al. 2019).

One person in the room could usually explain what those scores meant; however, caseworkers were not trained on the technical details during DCDA's initial deployment. Staff knowledge about "how the scores work" was therefore limited, with one caseworker simply stating that a high score implied that "there's a lot of risk factors." Others noted that the DCDA training materials reminded them of "graduate school, when you read papers and…only read the introduction, results, and discussion," admitted to not having read the training materials, and noted that they did not know what administrative records were used to compute risk scores or did not understand the meaning of those scores. Some risk scores were also unavailable for referrals that arrived after 5 pm, because the DCDA ran daily at the close of government business and did not retroactively compute scores for after-hours referrals during the next business day. RED team members still speculated about likely scores for those referrals but could not pull up the actual scores on their computer screens.

Results

The Challenge: Making Decisions with Ambiguous Information

We first examined the time allocated to different stages of the referral review process (Figure 2). On average, caseworkers spent 12.15 min per referral but could spend as little as 1 min and as much as 40 min (std. dev. = 5.57 min). We observed similar variation for different constituent parts of the referral reviews: reviews of priors (mean = 3.06 min, std. dev. = 2.86 min), reviews of current reports (mean = 4.51 min, std. dev. = 2.67 min), and initial deliberations (mean = 2.84 min, std. dev. = 3.49 min).

Panel (A) of Figure 2 shows that few referrals were discussed for less than 5 min. These referrals stood out for being unambiguous, allowing caseworkers to reach a decision quickly because neither parents nor children had prior system contact and because the allegations contained in the report pertained only to past behavior or remained clearly below the threshold for actionable maltreatment.

However, most referrals required caseworkers to reach a decision based on ambiguous information. Ambiguity resulted from: (1) incomplete information in maltreatment reports, as those who submitted those reports to CPS frequently did not know or did not mention key details about the alleged maltreatment, the affected child or children, or family contexts; (2) ongoing divorce proceedings or complex custody arrangements that made it difficult to judge the veracity of information given to CPS by a parent or family member; (3) long histories of familial contact with child welfare agencies or the criminal justice system, which presented challenges in identifying relevant historical patterns in light of the specific allegations contained in each report; and (4) discretionary threshold judgments, because child welfare caseworkers have substantial authority to establish whether specific allegations and evidence of familial hardship and dysfunction warrant a CPS investigation.

Against this backdrop of pervasive ambiguity, how caseworkers engaged with risk scores depended on the match—or mismatch—between their interpretation of priors and reports and the DCDA score that had been assigned to each referral.

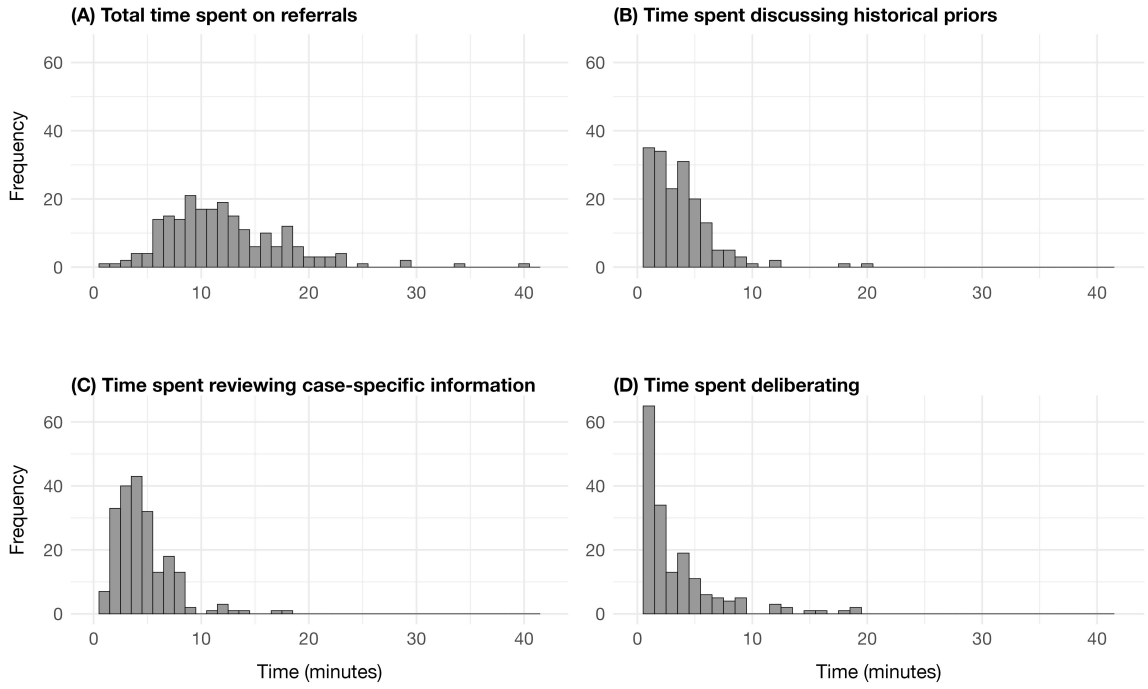

Figure 2: Time allocations across referrals.

Embracing Confirmatory Risk Scores: Alignment between Humans and DCDA

When caseworkers' interpretation of evidence aligned with risk scores—either because they strongly favored a screen-out decision and then encountered a low risk score, or because they strongly favored a screen-in decision and then encountered a high risk score—they simply acknowledged this alignment before proceeding to the next referral. In one case, caseworkers noted that the combination of a screen-out decision and a low risk score "makes sense" and "is nice." In other cases, they reasoned that a low score "sounds about right" and affirmed that their work was "done" when human assessments matched risk scores. According to one caseworker, "If your decision is supported, and it's not the opposite [. . .] you're good." We did not observe instances when caseworkers critiqued or questioned risk scores that aligned with their assessment of a referral.

Ignoring Mid-Range Risk Scores: Limited Information for Human Decision Making

DCDA generated two pieces of information: risk scores that were computed on a 20-point scale and the predicted out-of-home placement rate (expressed as a percentage value) at each point of that scale. However, caseworkers only discussed the numeric scores. Because rules for triaging maltreatment referrals did not allow caseworkers to defer their decisions, they needed to map those scores onto binary screen-in/screen-out decisions. This presented an interpretive challenge: Should

risk scores be treated as confirmatory if they fell into the top half (for screen-in decisions) or bottom half (for screen-out decisions) of the 20-point scale? Or should scores in the middle of the scale be discounted as weakly informative?

Caseworkers repeatedly expressed frustration with mid-range scores. For example, after reviewing one especially complex referral, the RED team was uncertain whether the evidence justified an investigation. When they consulted the risk score after their initial deliberations, which was an 8, they found it uninformative: "Mmmh-mmh. [. . . .] Yeah. Which is probably about how we're feeling. It's like, eh?" After reviewing another referral, a RED team member expressed a similar sentiment about a risk score of 12: "That's a very neutral number for me. I don't know what that means." Risk scores toward the middle of the 20-point scale were considered "useless" by a third caseworker, because they provided no clear direction that could inform and potentially improve binary screen-in/screen-out decisions in ambiguous cases.

Given the availability of a placement risk distribution (i.e., the predicted out-of-home placement rates associated with all possible scores), caseworkers could have decided on an acceptable threshold value, treating all scores associated with an above-threshold placement risk as favoring a screen-in decision. However, we did not observe this logic in the transcripts. This may be because (1) caseworkers ignored information about the predicted placement risks, focusing on a single numeric value that was prominently displayed on their computer screen, or because (2) predicted placement risks differed only moderately in the middle of the 20-point scale and may therefore not have been considered particularly meaningful.

Divergent Assessments: Rationalizing Disagreement between Humans and DCDA

Caseworkers often had specific expectations about the DCDA score they would encounter at the end of their deliberations, given each referral's characteristics. Sometimes they expressed those expectations by turning their engagement with DCDA into a guessing game: "What was the score on that one?"/"I'm gonna guess 8."/"That's a fair guess. Umm, 15."/"Do they have [. . .] history?"/"Mmh, no not too bad."/"Yeah, I'm gonna stick with 8." During those guessing games, RED team members consistently expected that extensive priors would cause DCDA to generate a high risk score. One caseworker opined that a child would likely be scored as being "at future risk of placement because she was in placement already," highlighting the assumed importance of prior CPS contact in DCDA computations. Other caseworkers noted that "I'm never usually surprised by [the referrals] that have a lot [of history]," "I know when [a referral] has 29 priors, [the score] is gonna be up there," and "I would imagine [the score] is pretty high" due to prior parental involvement with the criminal justice system. (The lasso regression model that powered DCDA analyzed historical data on CPS contact and public benefits access but not on parental criminal justice contact, whereas RED team members had access to "priors" from CPS records as well as parental court records. This subtle difference was largely lost during caseworker discussions.)

Conversely, caseworkers expected to encounter low risk scores for referrals without priors, because the DCDA score was then computed based on comparatively scarce data. When discussing a referral with a risk score of 1, they suggested that the score was lower than they had anticipated after reading the report "because there's probably no history."

Caseworkers' deepest engagement with risk scores occurred when DCDA scores were at odds with their expectations. They then attempted to rationalize a divergence between risk scores and their own assessment of a referral by speculating about the limitations and biases of DCDA, rather than re-evaluating their decisions. Their arguments fell into three categories.

Concerns about over-emphasizing historical data: Several caseworkers reasoned that the sensitivity of DCDA scores to historical patterns could yield artificially low scores for families without priors and inflated scores for families with extensive histories of state contact—regardless of the severity or immediacy of threats to children's health and wellbeing. An emphasis on historical data could also mask the importance of recent "triggering events" like parental divorce or a parent's decision to stop medication or counseling sessions. When discussing a referral with multiple criminal justice priors, one caseworker noted that "we would assign it" regardless of those priors (and regardless of the DCDA score) because a child had been physically injured during an altercation with their mother. However, this caseworker speculated that, without multiple priors, the score "would be a 1" because DCDA did not incorporate descriptions of the physical injury from the current referral.

In another instance, a caseworker reasoned that a risk score was "probably low" because the only historical data point was a single Driving Under the Influence (DUI) charge, whereas concerns about child wellbeing and safety were related to more recent family troubles (per the report, a divorce may have triggered parental alcohol and substance use problems). This caseworker suggested that risk scores could not "catch" those recent developments because they were computed using historical data rather than information from the current report. Conversely, caseworkers defended their decision to screen-out another referral despite a risk score of 19 by noting that existing priors and concerns about parental substance use "are from the past" and had already been investigated by CPS.

Concerns about over-emphasizing specific risk factors: Several caseworkers speculated that discrepancies between human screening decisions and machine-generated scores could be explained by the over- or under-weighting of specific data points. DCDA used lasso regularized regressions to compute risk scores, which shrink coefficients of less important predictors toward zero by applying a penalty term to ordinary least-squares regressions. This method facilitates feature selection in high-dimensional data and aims to improve interpretability, making predictive models more sensitive toward salient predictors while effectively eliminating others. RED team members did not display the requisite technical knowledge to evaluate this approach in detail, which was reasonable given the lack of training about DCDA's complex methodology. However, they still speculated that risk scores were particularly sensitive to parental alcohol abuse, allegations of sexual abuse, and abuse allegations involving young children.

In their own decisions, however, caseworkers tended to downweight parental DUI or public intoxication charges or reports of recent parental alcohol consumption (which were common) unless those could be tied to ongoing threats to a child's health and wellbeing. As one caseworker noted, regular parental drinking was not sufficient to trigger a CPS investigation "but if [the child] says they're driving her drunk, then that's [different]." In another case, a maltreatment report contained allegations of parental alcohol abuse but also noted that the parents had "been sober for two weeks." Caseworkers speculated that DCDA could not capture those subtle distinctions and specific sequences of events and thus generated inflated risk scores "probably because of the alcohol." They also agreed, in two separate instances, that DCDA should be specifically sensitive to child fatalities in a family. "With a fatality I don't know how [the score] can't be high," one caseworker noted, while another asserted that "if it's not high with a fatality then there's, there's a problem [with the tool]."

Concerns about missing contextual knowledge: Caseworkers' screening decisions were often holistic. They factored in historical priors as well as qualitative information from the maltreatment reports, which had not been fed into DCDA. Despite the wide range of datapoints used to compute DCDA scores—a total of 460 predictors—caseworkers also had the advantage of relying on their local knowledge and expertise, particularly in understanding policies and case nuances. This included knowledge about (1) the perceived trustworthiness of reporters (i.e., caseworkers trusted statements from healthcare workers or law enforcement officers but repeatedly questioned the veracity of statements from young children or from adults who were caught up in divorce proceedings); (2) the scope of non-CPS services that may be available to families to address at-home troubles, such as individual therapy, family counseling, or substance abuse programs; (3) the interplay of multiple contextual factors, such as substance abuse while breastfeeding, alcohol abuse while driving, or physical altercations while on a restraining order; (4) the complexity of custody arrangements that required children to spend time with each parent, to cohabitate temporarily with a parent's new partner, or to reside temporarily with grandparents or other relatives; and (5) common-sense understandings of what constitutes appropriate behavior for children at different ages and for parents who are facing different types of hardship (economic deprivation, mental health problems, family separation due to parental incarceration, or parental stress while caring for newborn or adolescent children).

In several instances, caseworkers explained discrepancies between their decisions and DCDA scores by asserting that the PRM tool lacked such contextual knowledge. For example, caseworkers justified four screen-out decisions by noting that families were already in contact with social workers or that parents had already been referred to a substance abuse counselor (under state regulations, caseworkers were permitted to address new concerns in an ongoing assessment, rather than opening a new one). In another instance, they suggested that a risk score was inflated because it factored in prior allegations against a biological parent (with whom the child did not reside overnight) rather than a new step parent (with whom the child temporarily co-habited, according to information from the report). And in a case involving suspected physical and supervisory neglect—which was screened

out—a caseworker argued that a risk score of 19 was inflated because it did not account for nuances in parenting: "I'm not hearing that [the child] is malnourished because he doesn't have any food [. . .] And he gets soap to wash himself, even if it's small pieces. And he eventually made it home from school that day."

We hypothesized that the tendency to rationalize divergences and retain the original (human) decision should prevent a clear association between risk scores and screening decisions. We tested this by regressing those screen-in/screen-out decisions on risk scores, finding no significant association ($\beta = -0.02$, SE $= 0.04$, p value $= 0.72$). We considered this as confirmatory evidence that DCDA scores did not lead caseworkers to reconsider their decisions, except in isolated instances. We found such instances to be rare. In one discussion, caseworkers worried that their screen-out decision could be overturned by a supervisor during a secondary review "because of the score." However, divergent risk scores did not usually prompt caseworkers to revise their decision. As one caseworker noted, "I'm surprised by the score [but] it doesn't really actually change my mind necessarily [. . .] For the most part it doesn't change what I think."

We found only two cases where a high risk score changed the final decision. In the first case, caseworkers had originally decided to screen out a referral that contained allegations of physical abuse but lacked crucial information. After seeing a risk score of 16, one caseworker urged their colleagues to revise the decision by taking the ambiguity of evidence seriously: "The only thing that is keeping me on the fence is the lacerations."/"But that could also be from scratching."/"That's what makes me on the fence is that they said that they can't rule it out." In the second case, which involved allegations of sexual abuse, caseworkers had originally decided to screen out the referral because the child already had another open referral. But after seeing a risk score of 20, one caseworker noted that "I don't feel comfortable" with the decision and opined that "somebody needs to [. . .] talk to the family." They then messaged colleagues who had been assigned to review the open referral, alerting them to newly reported concerns and proposing to merge two referrals into a single investigation. These two exceptional cases corroborate recent suggestive evidence of risk aversion among CPS caseworkers: When they had access to machine-generated scores, caseworkers were no less likely to screen out low-risk cases but were slightly more likely to screen in high-risk cases, possibly due to an elevated concern about false negative decisions (Fitzpatrick et al. 2025).

Discussion

Research on PRM tools has historically focused on measuring their predictive fairness and accuracy (Corbett-Davies et al. 2017; Daley et al. 2016; Duwe and Kim 2017; Flores, Bechtel, and Lowenkamp 2016; Imai et al. 2023; Montana et al. 2023; Obermeyer et al. 2019; Pastaltzidis et al. 2022; Pavlou et al. 2015; Zanger-Tishler, Nyarko, and Goel 2024) and has more recently focused on assessing their impact on administrative outcomes (Bhatt et al. 2024; Cheng et al. 2022; Eiermann 2024; Fitzpatrick et al. 2025; Grimon and Mills 2025; Imai et al. 2023; Kleinberg et al. 2018a; Parker et al. 2022; Stevenson 2018). Our study used uniquely suitable (and difficult-to-obtain) data to study human–machine interactions directly.

Understanding those interactions is important because the integration of PRM tools into administrative decision making—also known as "algorithm-in-the-loop" (Green and Chen 2019b)—can prompt a variety of responses. First, decision makers may *embrace* risk scores as a cost-effective innovation that can improve the targeting of state interventions (Kleinberg et al. 2018b). Second, they may *adjust* their decision making to account for perceived benefits and limitations of risk scores (Cheng et al. 2022). Third, they may *ignore* risk scores even as they remain formally compliant with policies that mandate or encourage their use (Stevenson and Doleac 2024). Fourth, they may *resist* the adoption of PRM tools, especially if those tools are perceived as threats to professional expertise and autonomy (Brayne and Christin 2021). The impacts of PRM tools—which can shape social and racial inequalities when an organization processes tens of thousands or even millions of cases, but which have repeatedly been found to be less than straightforward in practice (Eiermann 2024; Imai et al. 2023; Sloan et al. 2025; Stevenson 2018; Stevenson and Doleac 2024)—therefore depend not simply on the predictive accuracy of such tools but on the dynamics that ensue when human judgment and machine judgment are integrated into a single decision-making loop.

Our analyses of human–machine interactions showed that child welfare caseworkers closely followed protocol in consulting DCDA scores at the end of their deliberations, acknowledged the information value of scores at the high/low extremes of the distribution, and also defended (human) assessments against (machine) challenges. Specifically, this meant that caseworkers accepted particularly high and low scores that directly confirmed their decisions but ignored weakly informative scores in the middle of the risk distribution. When confronted with scores that diverged from their professional judgment, caseworkers rationalized those divergences by discussing the perceived limitations of DCDA. These findings show that the decision to embrace or ignore risk scores or to compensate for their perceived limitations depends both on the position of a specific score within the range of all possible scores (with greater attention paid at the extremes, and mid-range scores discounted) and also on the priors and perceptions that frontline officials have about PRM tools. Under the algorithm-in-the-loop framework that informed the introduction of DCDA into an existing process of structured CPS decision making, risk scores were not treated as decisive factors but as "discussion points" and outputs of an "experimental technology," leading to strategies that allowed caseworkers to sift through ambiguous evidence and complex family histories while also accounting for the perceived shortcomings of this technology (Preston-Shoot 2001, P. 9). Those strategies rarely led caseworkers to revise their decisions.

This dynamic can plausibly explain why the observed impacts of algorithmic tools are often uneven or unexpectedly modest even when those tools are adopted for routine use (Fitzpatrick et al. 2025; Stevenson 2018). Local compensatory strategies—that is, strategies that emerge as street-level administrators navigate formalized processes and the exigencies of their daily work during their interactions with machine-generated risk scores—can suppress the net impact of algorithmic tools on administrative decisions and lead to differential impacts across jurisdictions and domains of administration. More generally, our study highlights the importance

of studying new technologies of governance not just as "black boxes" with opaque inner workings (Winner 1993) or as engines of quantification that can decisively shape how institutional actors make consequential decisions (Mackenzie 2008) but as problem-solving tools that require buy-in from street-level professionals (Brayne and Christin 2021) and are expected to do particular and contextually specific kinds of work when they are embedded into larger systems of administration (Green and Chen 2019b). Whether they can—and will—do this work therefore depends as much on the mathematical architecture of the tools as it depends on the organizational architecture that supports their deployments.

In our study of Douglas County, we did not observe a straightforward deference to machine-generated scores, as might be expected if quantified indices of risk were perceived as particularly trustworthy (or as particularly useful shortcuts) at the street level (Porter 1995), if administrative decision making in the child welfare system was primarily characterized by "reactivity to indicators, as if working through a dashboard" (Fourcade and Gordon 2020, P. 87), or if caseworkers had to expect routine pushback from supervisors (Evans and Harris 2004). The well-established SDM framework may have discouraged such deference by providing caseworkers with a familiar, trusted, and officially sanctioned process that emphasized consensus-based decision making during local triage meetings. In contrast to some recent research on risk scoring in the criminal justice system (Brayne and Christin 2021), we also did not find evidence of overt resistance, nor did we find that caseworkers considered DCDA as a threat to their professional authority. This may be because our data collection occurred during an early phase of the PRM tool's deployment (and does not capture subsequent updates to local decision-making processes by DCDHS) or because the average tenure of child welfare caseworkers is comparatively short and their institutional position weak, relative to judges at parole board hearings.

A key question is therefore how human–machine interactions differ if (1) training procedures change and decision-makers' exposure to PRM tools increases, (2) policies evolve to encourage greater deference to machine-generated scores, especially in understaffed and underfunded welfare agencies, (3) algorithmic tools are deployed in contexts where decision making is less formalized and decisions must be made much quicker, and (4) newer AI tools analyze information that is often excluded from regression-based risk models but is considered pertinent by caseworkers, such as qualitative descriptions of maltreatment allegations and family contexts. Future research should directly engage with those scenarios.

References

Accominotti, Fabien. 2021. "The Aesthetics of Hierarchy." *The British Journal of Sociology* 72(2):196–202. https://doi.org/10.1111/1468-4446.12835.

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. 2024. *Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology*. Working Paper 31422. NBER Working Paper Series. Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w31422

Airoldi, Massimo. 2022. *Machine Habitus: Toward a Sociology of Algorithms*. Cambridge, MA: Polity Press.

Amoore, Louise. 2009. "Lines of Sight: On the Visualization of Unknown Futures." *Citizenship Studies* 13(1):17–30. https://doi.org/10.1080/13621020802586628

Beraldo, Davide and Stefania Milan. 2019. "From Data Politics to the Contentious Politics of Data." *Big Data & Society* 6(2):205395171988596. https://doi.org/10.1177/2053951719885967

Berk, Richard. 2017. "An Impact Assessment of Machine Learning Risk Forecasts on Parole Board Decisions and Recidivism." *Journal of Experimental Criminology* 13(2):193–216. https://doi.org/10.1007/s11292-017-9286-2

Bhatt, Monica P., Sara B. Heller, Max Kapustin, Marianne Bertrand, and Christopher Blattman. 2024. "Predicting and Preventing Gun Violence: An Experimental Evaluation of READI Chicago." *The Quarterly Journal of Economics* 139(1):1–56. https://doi.org/10.1093/qje/qjad031

Brayne, Sarah. 2017. "Big Data Surveillance: The Case of Policing." *American Sociological Review* 82(5):977–1008. https://doi.org/10.1177/0003122417725865

Brayne, Sarah. 2020. *Predict and Surveil: Data, Discretion, and the Future of Policing*. Oxford, UK: Oxford University Press

Brayne, Sarah and Angèle Christin. 2021. "Technologies of Crime Prediction: The Reception of Algorithms in Policing and Criminal Courts." *Social Problems* 68(3):608–24. https://doi.org/10.1093/socpro/spaa004

Burrell, Jenna and Marion Fourcade. 2021. "The Society of Algorithms." *Annual Review of Sociology* 47(1):213–37. https://doi.org/10.1146/annurev-soc-090820-020800

Burton, Jason W., Mari-Klara Stein, and Tina Blegind Jensen. 2020. "A Systematic Review of Algorithm Aversion in Augmented Decision Making." *Journal of Behavioral Decision Making* 33(2):220–39. https://doi.org/10.1002/bdm.2155

Cheney-Lippold, John. 2011. "A New Algorithmic Identity: Soft Biopolitics and the Modulation of Control." *Theory, Culture & Society* 28(6):164–81. https://doi.org/10.1177/0263276411424420

Cheng, Hao-Fei, Logan Stapleton, Anna Kawakami, Venkatesh Sivaraman, Yanghuidi Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. "How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions." Pp. 1–22 in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22*. New York, NY: ACM. https://doi.org/10.1145/3491102.3501831

Coglianese, Cary and Lavi M. Ben Dor. 2020. "AI in Adjudication and Administration." *Brooklyn Law Review* 86:791–838.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. "Algorithmic Decision Making and the Cost of Fairness." Pp. 797–806 in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Halifax, Canada: ACM. https://doi.org/10.1145/3097983.3098095

Cuccaro-Alamin, Stephanie, Regan Foust, Rhema Vaithianathan, and Emily Putnam-Hornstein. 2017. "Risk Assessment and Decision Making in Child Protective Services: Predictive Risk Modeling in Context." *Children and Youth Services Review* 79:291–98. https://doi.org/10.1016/j.childyouth.2017.06.027

Cuellar, Mariano-Florentino. 2016. "Artificial Intelligence and the Administrative State." *Regulatory Review in Depth* 5(2):19–30.

Cuellar, Mariano-Florentino. 2017. "A Simpler World: On Pruning Risks and Harvesting Fruits in a Orchard of Whispering Algorithms Symposium - Future-Proofing Law: From

RDNA to Robots." *U.C. Davis Law Review* 51(1):27–50. https://doi.org/10.2139/ssrn.3044720

Daley, Dyann, Michael Bachmann, Brittany A. Bachmann, Christian Pedigo, Minh-Thuy Bui, and Jamye Coffman. 2016. "Risk Terrain Modeling Predicts Child Maltreatment." *Child Abuse & Neglect* 62:29–38. https://doi.org/10.1016/j.chiabu.2016.09.014

Desrosières, Alain. 2002. *The Politics of Large Numbers: A History of Statistical Reasoning*. Cambridge, MA: Harvard University Press.

Duwe, Grant and KiDeuk Kim. 2017. "Out With the Old and in With the New? An Empirical Comparison of Supervised Learning Algorithms to Predict Recidivism." *Criminal Justice Policy Review* 28(6):570–600. https://doi.org/10.1177/0887403415604899

Edwards, Frank and Christopher Wildeman. 2018. "Characteristics of the Front-Line Child Welfare Workforce." *Children and Youth Services Review* 89:13–26. https://doi.org/10.1016/j.childyouth.2018.04.013

Eiermann, Martin. 2024. "Algorithmic Risk Scoring and Welfare State Contact Among US Children." *Sociological Science* 11:707–42. https://doi.org/10.15195/v11.a26

Ellis, Kathryn, Ann Davis, and Kirstein Rummery. 1999. "Needs Assessment, Street-Level Bureaucracy and the New Community Care." *Social Policy & Administration* 33(3):262–80. https://doi.org/10.1111/1467-9515.00150

Espeland, Wendy Nelson and Mitchell L. Stevens. 1998. "Commensuration as a Social Process." *Annual Review of Sociology* 24:313–43. https://doi.org/10.1146/annurev.soc.24.1.313

Espeland, Wendy Nelson and Mitchell L. Stevens. 2008. "A Sociology of Quantification." *European Journal of Sociology/Archives Européennes de Sociologie* 49(3):401–36. https://doi.org/10.1017/S0003975609000150

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York, NY: St. Martin's Press.

Evans, Tony and John Harris. 2004. "Street-Level Bureaucracy, Social Work and the (Exaggerated) Death of Discretion." *British Journal of Social Work* 34(6):871–95. https://doi.org/10.1093/bjsw/bch106

Fitzpatrick, Maria D., Katharine Sadowski, and Christopher Wildeman. 2025. "Algorithms and Decision-Making: Evidence from Child Maltreatment Reports." *Journal of Human Resources* 0224-13437R2. https://doi.org/10.3368/jhr.0224-13437R2

Flores, Anthony W., Kristin Bechtel, and Christopher T. Lowenkamp. 2016. "False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks." *Federal Probation* 80:38.

Fourcade, Marion. 2021. "Ordinal Citizenship." *The British Journal of Sociology* 72(2):154–73. https://doi.org/10.1111/1468-4446.12839

Fourcade, Marion and Jeffrey Gordon. 2020. "Learning Like a State: Statecraft in the Digital Age." *Journal of Law and Political Economy* 1(1):78–108. https://doi.org/10.5070/LP61150258

Fourcade, Marion and Kieran Healy. 2024. *The Ordinal Society*. Cambridge, MA: Harvard University Press. https://doi.org/10.4159/9780674296688

Green, Ben and Yiling Chen. 2019a. "Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments." Pp. 90–99 in *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT '19*. New York, NY: ACM. https://doi.org/10.1145/3287560.3287563

Green, Ben and Yiling Chen. 2019b. "The Principles and Limits of Algorithm-in-the-Loop Decision Making." Pp. 1–24 in *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW). New York, NY: ACM. https://doi.org/10.1145/3359152

Green, Ben and Yiling Chen. 2021. "Algorithmic Risk Assessments Can Alter Human Decision-Making Processes in High-Stakes Government Contexts." Pp. 1–33 in *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2). New York, NY: ACM. https://doi.org/10.1145/3479562

Grimon, Marie-Pascale and Christopher Mills. 2025. "Better Together? A Field Experiment on Human-Algorithm Interaction in Child Protection." arXiv:2502.08501.

Imai, Kosuke, Zhichao Jiang, D. James Greiner, Ryan Halen, and Sooahn Shin. 2023. "Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment." *Journal of the Royal Statistical Society Series A: Statistics in Society* 186(2):167–89. https://doi.org/10.1093/jrsssa/qnad010

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018a. "Human Decisions and Machine Predictions." *The Quarterly Journal of Economics* 133(1):237–93. https://doi.org/10.1093/qje/qjx032

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Cass R. Sunstein. 2018b. "Discrimination in the Age of Algorithms." *Journal of Legal Analysis* 10(2005):113–74. https://doi.org/10.1093/jla/laz001

Koopman, Colin. 2019. *How We Became Our Data: A Genealogy of the Informational Person*. Chicago, IL: University of Chicago Press. https://doi.org/10.7208/chicago/9780226626611.001.0001

Lin, Zhiyuan "Jerry", Jongbin Jung, Sharad Goel, and Jennifer Skeem. 2020. "The Limits of Human Predictions of Recidivism." *Science Advances* 6(7):eaaz0652. https://doi.org/10.1126/sciadv.aaz0652.

Lipsky, Michael. 2010. *Street-Level Bureaucracy: Dilemmas of the Individual in Public Service*. New York, NY: Russell Sage Foundation.

Mackenzie, Donald. 2008. *An Engine, Not a Camera How Financial Models Shape Markets*. Cambridge, MA: MIT Press.

Mahmud, Hasan, A. K. M. Najmul Islam, Syed Ishtiaque Ahmed, and Kari Smolander. 2022. "What Influences Algorithmic Decision-Making? A Systematic Literature Review on Algorithm Aversion." *Technological Forecasting and Social Change* 175:121390. https://doi.org/10.1016/j.techfore.2021.121390

Mau, Steffen. 2019. *The Metric Society: On the Quantification of the Social*. New York, NY: John Wiley & Sons.

Maynard-Moody, Steven and Michael Musheno. 2012. "Social Equities and Inequities in Practice: Street-Level Workers as Agents and Pragmatists." *Public Administration Review* 72(s1):S16–23. https://doi.org/10.1111/j.1540-6210.2012.02633.x

Merry, Sally Engle. 2016. *The Seductions of Quantification: Measuring Human Rights, Gender Violence, and Sex Trafficking*. Chicago, IL: University of Chicago Press. https://doi.org/10.7208/chicago/9780226261317.001.0001

Montana, Erika, Daniel S. Nagin, Roland Neil, and Robert J. Sampson. 2023. "Cohort Bias in Predictive Risk Assessments of Future Criminal Justice System Involvement." *Proceedings of the National Academy of Sciences* 120(23):e2301990120. https://doi.org/10.1073/pnas.2301990120

Morgan, Kimberly J. and Ann Shola Orloff. 2017. *The Many Hands of the State: Theorizing Political Authority and Social Control*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/9781316471586

Mullainathan, Sendhil and Ziad Obermeyer. 2022. "Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care." *The Quarterly Journal of Economics* 137(2):679–727. https://doi.org/10.1093/qje/qjab046

Neil, Roland and Michael Zanger-Tishler. 2025. "Algorithmic Bias in Criminal Risk Assessment: The Consequences of Racial Differences in Arrest as a Measure of Crime." *Annual Review of Criminology* 8:97–119. https://doi.org/10.1146/annurev-criminol-022422-125019

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366(6464):447–53. https://doi.org/10.1126/science.aax2342

Parker, Elizabeth M., Jason R. Williams, Peter J. Pecora, and Daniel Despard. 2022. "Examining the Effects of the Eckerd Rapid Safety Feedback Process on the Occurrence of Repeat Maltreatment among Children Involved in the Child Welfare System." *Child Abuse & Neglect* 133:105856. https://doi.org/10.1016/j.chiabu.2022.105856

Pastaltzidis, Ioannis, Nikolaos Dimitriou, Katherine Quezada-Tavarez, Stergios Aidinlis, Thomas Marquenie, Agata Gurzawska, and Dimitrios Tzovaras. 2022. "Data Augmentation for Fairness-Aware Machine Learning: Preventing Algorithmic Bias in Law Enforcement Systems." Pp. 2302–14 in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22. New York, NY: ACM.

Pavlou, Menelaos, Gareth Ambler, Shaun R. Seaman, Oliver Guttmann, Perry Elliott, Michael King, and Rumana Z. Omar. 2015. "How to Develop a More Accurate Risk Prediction Model When There Are Few Events." *BMJ* 351:h3868. https://doi.org/10.1136/bmj.h3868

Porter, Theodore. 1995. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, NJ: Princeton University Press. https://doi.org/10.1515/9781400821617

Preston-Shoot, Michael. 2001. "Regulating the Road of Good Intentions: Observations on the Relationship between Policy, Regulations and Practice in Social Work." *Practice* 13(4):5–20. https://doi.org/10.1080/09503150108411523

Pruss, Dasha. 2023. "Ghosting the Machine: Judicial Resistance to a Recidivism Risk Assessment Instrument." Pp. 312–23 in *2023 ACM Conference on Fairness, Accountability, and Transparency*. Chicago, IL: ACM. https://doi.org/10.1145/3593013.3593999

Saeed, Subha, Rahul Patel, and Rachel Odeyemi. 2022. "Calibrating Readmission Risk Prediction Models for Determining Post-Discharge Follow-up Timing." *Journal of Community Hospital Internal Medicine Perspectives* 12(4):24–28. https://doi.org/10.55729/2000-9666.1036

Saxena, Devansh, Karla Badillo-Urquiola, Pamela J. Wisniewski, and Shion Guha. 2020. "A Human-Centered Review of Algorithms Used within the U.S. Child Welfare System." Pp. 1–15 in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Honolulu, HI: ACM. https://doi.org/10.1145/3313831.3376229

Scott, James C. 1999. *Seeing like a State: How Certain Schemes to Improve the Human Condition Have Failed*. New Haven, CT: Yale University Press.

Seaver, Nick. 2018. "What Should an Anthropology of Algorithms Do?" *Cultural Anthropology* 33(3):375–85. https://doi.org/10.14506/ca33.3.04

Skeem, Jennifer, Nicholas Scurich, and John Monahan. 2020. "Impact of Risk Assessment on Judges' Fairness in Sentencing Relatively Poor Defendants." *Law and Human Behavior* 44(1):51–9. https://doi.org/10.1037/lhb0000360

Sloan, CarlyWill, George Naufal, and Heather Caspers. 2025. "The Effect of Risk Assessment Scores on Judicial Behavior and Defendant Outcomes." *Journal of Human Resources* 60(5):1778–810. https://doi.org/10.3368/jhr.0221-11470R3

Stevenson, Megan. 2018. "Assessing Risk Assessment in Action." *Minnesota Law Review* 103:303. https://doi.org/10.2139/ssrn.3016088

Stevenson, Megan T. and Jennifer L. Doleac. 2024. "Algorithmic Risk Assessment in the Hands of Humans." *American Economic Journal: Economic Policy* 16(4):382–414.

U.S. Department of Health & Human Services, Administration for Children and Families, Administration on Children, Youth and Families, Children's Bureau. 2024. *Child Maltreatment 2022*.

Vaithianathan, Rhema, Haley Dinh, Allon Kalisher, Chamari Kithulgoda, Emily Kulick, Megh Mayur, Athena Ning, Diana Benavides Prado, and Emily Putnam-Hornstein. 2019. *Implementing a Child Welfare Decision Aide in Douglas County: Methodology Report*. Auckland, New Zealand: Centre for Social Data Analytics.

Weber, Max. 1978. *Economy and Society: An Outline of Interpretive Sociology*. Berkeley, CA: University of California Press.

Winner, Langdon. 1993. "Upon Opening the Black Box and Finding It Empty: Social Constructivism and the Philosophy of Technology." *Science, Technology, & Human Values* 18(3):362–78. https://doi.org/10.1177/016224399301800306

Zanger-Tishler, Michael, Julian Nyarko, and Sharad Goel. 2024. "Risk Scores, Label Bias, and Everything but the Kitchen Sink." *Science Advances* 10(13):eadi8411. https://doi.org/10.1126/sciadv.adi8411

Acknowledgments: The authors are grateful to Ruby Richards and Nicole Adams for feedback on earlier drafts of this manuscript and the Douglas County Department of Human Services for providing data throughout this project.

Martin Eiermann: Department of Sociology, University of Wisconsin-Madison.
E-mail: meiermann@wisc.edu.
Maria Fitzpatrick: Books School of Public Policy, Cornell University; National Bureau of Economic Research. E-mail: maria.d.fitzpatrick@cornell.edu.
Katharine Sadowski: Graduate School of Education, Stanford University.
E-mail: ksadow@stanford.edu.
Christopher Wildeman: Department of Sociology, Duke University; Sanford School of Public Policy, Duke University; ROCKWOOL Foundation Research Unit.
E-mail: christopher.wildeman@duke.edu.