



## Policy Analysis

# Algorithmic Bias in Predictive Models for Child Protective Services

Chiara Andersson<sup>1\*</sup>, Diego Dvorak<sup>2</sup> and Claire Larsen<sup>3</sup>

<sup>1</sup>Department of Bioethics, Bohemian University of Health Policy and Ethics, Czech Republic

<sup>2</sup>Department of Bioethics, Jutland University Centre for Bioethics, Denmark

<sup>3</sup>Department of Public Health Ethics, Jutland University Centre for Bioethics, Denmark

### ABSTRACT

*Predictive risk models such as the Allegheny Family Screening Tool are increasingly used to help child protective services screeners decide which referrals warrant investigation, on the premise that a statistical score can outperform inconsistent human judgment under severe time pressure.*

*This policy analysis reviews the ethical and empirical literature on how these tools are built and validated, the specific mechanisms by which they can encode racial, disability, and poverty-linked disparities, and what happens when their scores are handed to human decision-makers rather than acting alone.*

*The evidence indicates that variables correlated with race, disability, and poverty function as proxies for risk rather than direct measures of it, that human override of algorithmic scores has itself been shown to introduce racial disparity rather than reliably correct it, and that most deployed tools have not been evaluated against fairness-reporting standards that would make these problems visible before deployment.*

*We examine governance proposals converging on independent pre- and post-deployment fairness auditing, deliberate scrutiny of proxy variables, genuine institutional support for caseworker discretion, and mechanisms for family transparency and contestability.*

*We conclude that no currently deployed system satisfies all of these conditions, and that the field's own reviewers still describe the underlying validation problem, testing a model against outcomes that a historically biased system itself produced, as unresolved.*

**Keywords:** algorithmic bias; child protective services; predictive risk modeling; racial disparity; data justice

**Citation:** Andersson C, Dvorak D and Larsen C (2026) Algorithmic Bias in Predictive Models for Child Protective Services. *Bioethics*. 1(2):01-06.

**Corresponding Author:** Chiara Andersson

**E-mail Address:** c.andersson@buhpe.cz

**Article History:** Received 12 June 2026 | Accepted 18 September 2026 | Published 15 December 2026

**Copyright:** © 2026 Chiara Andersson, et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (<https://creativecommons.org/licenses/by/4.0/>).

## INTRODUCTION

Child protective services agencies across the United

States, New Zealand, the United Kingdom, and elsewhere have spent the last decade experimenting with predictive risk models meant to help overworked

hotline screeners decide which referrals warrant a full investigation. The best-known of these, the Allegheny Family Screening Tool (AFST) deployed in Allegheny County, Pennsylvania, generates a numeric risk score from administrative records -- jail bookings, behavioral health service use, prior child welfare contacts -- at the moment a call comes in [1]. The pitch behind these systems is straightforward: caseworkers are drowning in referrals, human judgment is inconsistent and vulnerable to bias of its own, and a model trained on historical outcomes can in principle flag the highest-risk cases more reliably than an intake worker skimming a file in a few minutes [2]. Early validation work on the AFST found that children it identified as high-risk did go on to have more injury-related hospital encounters than lower-scored children, which its developers took as evidence that the tool was tracking something real about future harm rather than simply reproducing noise [1].

The trouble is that "something real about future harm" is not the same as "something fair, transparent, and ethically defensible to act on." Child protective services predictive models sit at an unusually fraught intersection: the data they are trained on record only what the existing system happened to notice and investigate, the populations most heavily represented in that data are disproportionately poor and disproportionately Black, and the decisions the models inform -- whether to open an investigation, whether to remove a child from a home -- carry consequences that are difficult to reverse and that fall unevenly across families to begin with [3]. This review surveys the ethical literature on algorithmic bias in child protection risk prediction: how these tools work and how they have been validated, the specific mechanisms by which they can encode and amplify existing disparities, what happens when they interact with human decision-makers rather than replacing them, and what the emerging governance proposals actually ask agencies to do differently. Table 1 summarizes four commonly voiced ethical concerns against what the empirical literature reviewed here has actually found, and Figure 1 offers a schematic illustration of how transparency, contestability, and disparity risk vary across three points on a spectrum of predictive-model designs used or proposed for child welfare screening.

### HOW PREDICTIVE RISK MODELS REACH THE SCREENING DECISION

Predictive risk models in this domain are not black boxes in the deep-learning sense; the AFST and most of its analogues are built on relatively interpretable statistical methods -- typically logistic regression or gradient-boosted trees -- trained on linked administrative data that a call screener would not otherwise be able to review in the few minutes available before deciding whether to open a case [2]. This is a continuation of a much older tradition of

actuarial risk assessment in child welfare, which struggled with its own reliability and predictive-validity problems long before machine learning entered the picture and which established, well before the current debate, that a tool's statistical accuracy on a validation sample does not by itself settle whether it should guide a specific family's case [4]. The input variables are proxies rather than direct measures of harm: prior child welfare involvement, public benefits receipt, jail or behavioral health system contact, and neighborhood-level indicators all correlate with future substantiated maltreatment in the historical record, but each of them is also a marker of poverty, disability, or prior surveillance rather than of parenting quality as such [5]. A detailed audit of the AFST's variable construction found that several inputs function, in effect, as proxies for the same underlying disadvantage measured multiple times, inflating a family's score for being poor and previously investigated rather than for anything specific to the referral at hand [6].

This is where the ethical literature parts ways with a narrower, purely statistical framing of "bias" as calibration error. A model can be well-calibrated -- its predicted risk scores can track observed outcomes accurately within each demographic subgroup -- and still be ethically troubling if the outcome it is calibrated against is itself a product of biased prior decision-making, a problem the fairness and machine-learning literature has long recognized applies well beyond this one domain [7]. In child welfare, the outcome variable most models are trained on is not "maltreatment" but "substantiated report," and substantiation itself reflects which families get reported, investigated, and formally confirmed under a system that has historically over-surveilled Black and low-income households relative to their actual rates of child harm [8]. A recent systematic review of machine-learning models used in child protection prediction found that fairness auditing and methodological reporting standards (such as TRIPOD and PROBAST) are inconsistently applied across the published tools, making it difficult to know, tool by tool, whether disparities identified after deployment were foreseeable at the design stage [9]. A companion review focused specifically on model calibration and reporting reached a similar conclusion: many maltreatment-risk prediction models have not been evaluated against the reporting standards that would let an outside agency assess their fairness properties before adoption [10].

### EVIDENCE ON RACIAL AND DISABILITY DISPARITIES

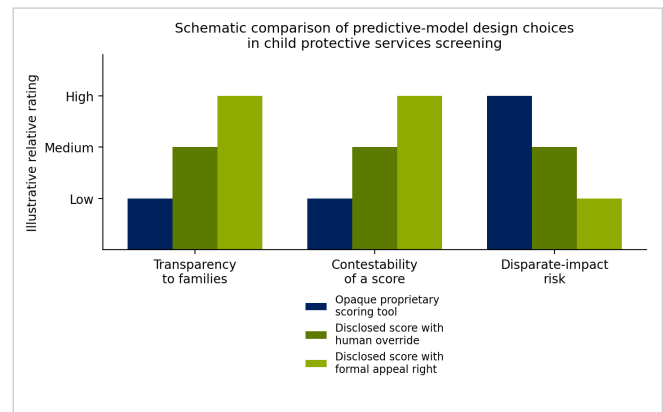
Because the AFST is the most extensively studied deployed tool, most of the direct empirical evidence on disparate impact in this literature concerns it specifically, though the underlying mechanisms are not unique to Allegheny County. A study examining the

tool's operation over several years found that its risk scores, and the screening decisions that followed from them, differed by race in ways not fully explained by underlying differences in maltreatment risk, and that human screeners' departures from the model's recommendation -- overriding a low score upward or a high score downward -- introduced a further, separate source of racial disparity on top of whatever the algorithm itself produced [11]. That finding matters because it complicates a common defense of these tools, namely that any bias identified is a data problem to be fixed rather than an argument against algorithmic screening altogether: if caseworkers' interactions with the tool's output introduce their own disparities, correcting the model's calibration alone will not resolve the problem observed in practice.

**Table 1. Commonly voiced ethical concerns about predictive risk models in child protective services, compared with what the empirical literature reviewed here has found.**

| Concern raised                   | What proponents argue                                                                                   | What the empirical/ethical literature finds                                                                                                                                     |
|----------------------------------|---------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Racial disparity                 | Disparities reflect real differences in family circumstances captured by the data, not the model itself | Risk scores and caseworker responses to them show race-linked disparities not fully explained by underlying risk, with human overrides adding a further layer of disparity [11] |
| Disability status as risk factor | Behavioral-health contact is a legitimate predictor of household stress                                 | Disability-correlated variables raise scores independent of documented harm, effectively treating disability itself as risk [12]                                                |
| Human oversight as safeguard     | A caseworker retains final authority, so the algorithm only advises                                     | Override patterns have been shown to introduce disparities rather than reliably correct the model's [11]                                                                        |
| Model accuracy as validation     | Calibration against outcomes shows the tool tracks real risk                                            | Reporting-standard reviews find inconsistent fairness auditing, and the outcome label itself reflects prior, potentially biased substantiation decisions [9,20]                 |

**Figure 1. Schematic, illustrative comparison of transparency, contestability, and disparate-impact risk across three points on the spectrum of predictive-model designs discussed in this review (relative Low/Medium/High ratings for illustration only, not derived from a single empirical study).**



A separate strand of this literature has examined disability status rather than race. An evaluation of the AFST's variable set found that inputs correlated with parental or child disability -- behavioral health service utilization chief among them -- systematically raised risk scores for disabled parents and their children independent of any documented harm, echoing longstanding concern in disability advocacy that child welfare systems treat disability itself as a risk factor rather than attending to whether adequate support was available to the family [12]. Table 1 places this alongside the racial-disparity findings because the underlying mechanism is structurally similar in both cases: a variable that correlates with a protected or vulnerable status gets folded into a risk score as though it were a direct signal of danger to a child, and the model has no way of distinguishing "this family faces structural disadvantage" from "this family poses elevated risk to a child."

Racial disproportionality in child welfare decision-making did not originate with predictive analytics, of course; disparity indices tracking Black children's overrepresentation in investigation, substantiation, and removal predate the introduction of any algorithmic tool by decades [8], and county-level qualitative work has documented practice interventions that measurably reduced disparate placement outcomes without any predictive model involved [13]. What the algorithmic-bias literature adds to that older disproportionality literature is a mechanism for how disparities already present in agency records get formalized, scaled, and made harder to see: a caseworker's individual bias is idiosyncratic and, in principle, correctable case by case, while a biased input variable baked into a model's weights is applied uniformly and silently across every referral the tool touches [14].

## TRANSPARENCY, CONTESTABILITY, AND THE LIMITS OF HUMAN OVERSIGHT

Most jurisdictions that have deployed predictive

screening tools have paired them with an explicit assurance that the algorithm advises rather than decides -- a human caseworker retains final authority over whether to open a case. Advocates and critics alike have identified reasons to doubt how much protection that pairing actually provides. Ethnographic and policy analysis of predictive risk modeling in New Zealand's child protection context argued that risk scores function less as neutral decision support and more as a mechanism that reframes structurally produced poverty as individual parental risk, shifting the political character of the intervention even when a human remains formally in charge of the final call [15]. Later work by the same author extended this into a broader account of decision variability in child protection, arguing that the ethical stakes of algorithmic tools cannot be separated from the wider question of how much interpretive discretion, and how much respect for family circumstance, the system is willing to preserve once a numeric score is placed in front of a decision-maker [16].

A parallel line of legal and policy scholarship has focused on due-process concerns: families are rarely told that a risk score influenced the decision to investigate them, have no established mechanism to see or contest the score, and the proprietary or administratively opaque nature of many of these systems makes independent audit difficult even for researchers, let alone for an affected parent [17]. A recent proposal for a legal-ethical framework governing algorithmic risk tools in child welfare, developed in the Spanish context but applicable more broadly, argues that meaningful contestability requires disclosure of a score's existence to the family, an accessible explanation of the factors driving it, and a formal channel for challenging it before consequential decisions are finalized -- protections that essentially no currently deployed child welfare risk tool provides in full [18]. Comparative work on predictive risk modeling as a signal of adversity has argued that these tools work best, ethically, when treated explicitly as adversity-detection instruments meant to trigger an offer of support, rather than as risk-of-maltreatment instruments meant to justify investigation or removal -- a reframing that would change substantially which variables belong in the model and what a high score should be understood to mean [19].

Substantiation decision-making research adds a further complication: even the ground-truth label these models are trained on is itself the product of variable, discretionary human judgment that differs across investigators and jurisdictions, which means a model trained to predict "substantiation" is, in an important sense, trained to predict the idiosyncrasies of the substantiation process rather than maltreatment as an independent fact [20]. Encyclopedia-level surveys of predictive tools in child welfare have summarized this as an unresolved methodological tension rather than a solved problem: no widely deployed tool has fully

escaped the circularity of being validated against outcomes that the existing, potentially biased system itself produced [21].

## TOWARD ACCOUNTABLE DESIGN AND GOVERNANCE

The reform proposals that have emerged from this literature converge on a smaller number of principles than the range of critiques might suggest. First, fairness auditing needs to happen before deployment and be repeated after it, using reporting standards that let outside researchers and affected communities assess a tool's disparate-impact properties rather than taking a developer's internal validation at face value [9,10]. Second, variable selection deserves scrutiny specifically for proxy effects -- inputs that function as stand-ins for race, disability, or poverty status should be flagged and either removed or explicitly justified against a demonstrated, independent relationship to child safety, not merely to substantiation history [6]. Third, meaningful human oversight requires more than formal authority to override a score; it requires training, workload, and institutional incentives that actually support caseworkers in exercising independent judgment rather than defaulting to the number in front of them, since evidence so far suggests that override patterns can introduce disparities of their own rather than correcting the model's [11].

Fourth, and perhaps most contested, is the question of whether families should be told that a score exists and given some route to contest it. The strongest version of this argument holds that a decision with removal or prolonged state supervision as a possible outcome cannot be ethically defensible without some form of notice and challenge, however administratively inconvenient that would be for high-volume screening operations [17,18]. A weaker but more institutionally realistic version calls for aggregate, rather than individual, transparency: regular public reporting of a tool's disparate-impact statistics by race, disability status, and poverty indicators, sufficient for community oversight bodies and researchers to hold an agency accountable even without giving individual families a right to see their own score [3]. Neither version has been adopted in full by any jurisdiction currently operating one of these tools, which is itself informative about how far institutional practice still lags the ethical literature surrounding it.

## CONCLUSION

Predictive risk models in child protective services were introduced to solve a real problem -- overloaded screeners making rapid, high-stakes decisions on incomplete information -- and the evidence that some of these tools track genuine differences in downstream harm is not negligible [1]. But the ethical literature surveyed here makes clear that statistical validity is not the same thing as fairness, and that a model calibrated against a historically biased substantiation

record can reproduce and formalize that bias at scale even while appearing accurate on its own terms [7,8]. The clearest findings concern proxy variables that encode race, disability, and poverty rather than risk as such [6,12], and the more unsettling finding is that human oversight, the safeguard most jurisdictions point to, has itself been shown to introduce disparities rather than reliably correcting for them [11]. Reform proposals converge on independent pre- and post-deployment fairness auditing, deliberate scrutiny of proxy variables, genuine support for caseworker discretion, and some meaningful form of transparency toward the families whose lives these scores shape, but no deployed system yet satisfies all of these conditions at once, and the field's own reviewers describe the underlying circularity -- validating a model against outcomes the existing system produced -- as still unresolved [21].

## DECLARATIONS

### Acknowledgements

The authors are grateful to colleagues at their respective departments of bioethics and public health ethics for helpful conversations that shaped the scope and framing of this review.

### Funding

This review received no external funding.

### Conflict of Interest

The authors declare no conflicts of interest relevant to this work.

### Author Contributions

Chiara Andersson conceived the review, conducted the initial literature synthesis, and drafted the manuscript. Diego Dvorak contributed the analysis of racial and disability disparity evidence and critically revised the manuscript for intellectual content. Claire Larsen contributed the sections on transparency, contestability, and governance proposals and approved the final version for submission. All authors read and approved the final manuscript.

### Ethics Statement

As a literature review synthesizing previously published, publicly available research and policy analysis, this article did not involve new data collection from human participants and did not require institutional ethics committee approval.

### Data Availability

No new data were generated or analyzed for this review. All sources discussed are cited in the reference list and are publicly available through their respective publishers.

## REFERENCES

1. Vaithianathan R, Putnam-Hornstein E, Chouldechova A, Benavides-Prado D, Berger R (2020) Hospital injury encounters of children identified by a predictive risk model for screening child maltreatment referrals. *JAMA Pediatrics*, DOI:[10.1001/jamapediatrics.2020.2770](https://doi.org/10.1001/jamapediatrics.2020.2770).
2. Cuccaro-Alamin S, Foust R, Vaithianathan R, Putnam-Hornstein E (2017) Risk assessment and decision making in child protective services: predictive risk modeling in context. *Children and Youth Services Review*, DOI:[10.1016/j.childyouth.2017.06.027](https://doi.org/10.1016/j.childyouth.2017.06.027).
3. Keddell E (2020) Case for an inequalities perspective in child protection. *Policy Quarterly*, DOI:[10.26686/pq.v16i1.6455](https://doi.org/10.26686/pq.v16i1.6455).
4. Camasso MJ, Jagannathan R (2000) Modeling the reliability and predictive validity of risk assessment in child protective services. *Children and Youth Services Review*, DOI:[10.1016/s0190-7409\(00\)00121-3](https://doi.org/10.1016/s0190-7409(00)00121-3).
5. Russell J (2015) Predictive analytics and child protection: constraints and opportunities. *Child Abuse & Neglect*, DOI:[10.1016/j.chiabu.2015.05.022](https://doi.org/10.1016/j.chiabu.2015.05.022).
6. Gerchick M, Jegede T, Shah T, Gutierrez A, Beiers S, Shemtov N, Xu K, Samant A, Horowitz A (2023) The devil is in the details: interrogating values embedded in the Allegheny Family Screening Tool. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, DOI:[10.1145/3593013.3594081](https://doi.org/10.1145/3593013.3594081).
7. Hutchinson B, Mitchell M (2019) 50 years of test (un)fairness: lessons for machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, DOI:[10.1145/3287560.3287600](https://doi.org/10.1145/3287560.3287600).
8. Kim H, Chenot D, Ji J (2011) Racial/ethnic disparity in child welfare systems: a longitudinal study utilizing the Disparity Index (DI). *Children and Youth Services Review*, DOI:[10.1016/j.childyouth.2011.02.021](https://doi.org/10.1016/j.childyouth.2011.02.021).
9. Bull C, Kisely S, Betts K, Hu Y (2025) Understanding the development, performance, fairness, and transparency of machine learning models used in child protection prediction: a systematic review. *Child Abuse & Neglect*, DOI:[10.1016/j.chiabu.2025.107630](https://doi.org/10.1016/j.chiabu.2025.107630).
10. Ahmed RS, Shaban M (2026) Prediction models for maltreatment risk: TRIPOD/PROBAST compliance, calibration, and fairness -- a systematic review. *Child Abuse & Neglect*, DOI:[10.1016/j.chiabu.2026.107923](https://doi.org/10.1016/j.chiabu.2026.107923).
11. Rittenhouse K, Putnam-Hornstein E, Vaithianathan R (2026) Algorithms, humans, and racial disparities in child protection systems: evidence from the Allegheny Family Screening Tool. *Journal of Policy Analysis and Management*, DOI:[10.1002/pam.70121](https://doi.org/10.1002/pam.70121).

12. Moura I, Powell RM, Kaye HS (2026) Evaluating disability bias in the Allegheny Family Screening Tool: implications for child welfare decisions. *Journal of Public Child Welfare*, DOI:[10.1080/15548732.2026.2689958](https://doi.org/10.1080/15548732.2026.2689958).
13. Pryce J, Lee W, Crowe E, Park D, McCarthy M, Owens G (2018) A case study in public child welfare: county-level practices that address racial disparity in foster care placement. *Journal of Public Child Welfare*, DOI:[10.1080/15548732.2018.1467354](https://doi.org/10.1080/15548732.2018.1467354).
14. Weber K, Morrison M (2020) The institutional analysis: a tool for diagnosing structural contributors to racial disproportionality and disparity in child welfare. In: *Racial Disproportionality and Disparities in the Child Welfare System*, DOI:[10.1007/978-3-030-54314-3\\_19](https://doi.org/10.1007/978-3-030-54314-3_19).
15. Keddell E (2015) The ethics of predictive risk modelling in the Aotearoa/New Zealand child welfare context: child abuse prevention or neo-liberal tool? *Critical Social Policy*, DOI:[10.1177/0261018314543224](https://doi.org/10.1177/0261018314543224).
16. Keddell E (2023) On decision variability in child protection: respect, interactive universalism and ethics of care. *Ethics and Social Welfare*, DOI:[10.1080/17496535.2022.2073381](https://doi.org/10.1080/17496535.2022.2073381).
17. Glaberson S (2023) Coding over the cracks: predictive analytics and child protection. SSRN, DOI:[10.2139/ssrn.4448913](https://doi.org/10.2139/ssrn.4448913).
18. Mugamba E (2025) Predictive protection or profiling? A legal-ethical framework for algorithmic risk tools in child welfare systems in Spain. *Research Square*, DOI:[10.21203/rs.3.rs-7388305/v1](https://doi.org/10.21203/rs.3.rs-7388305/v1).
19. de Haan I, Connolly M (2017) Anticipating risk: predictive risk modelling as a signal of adversity. In: *Beyond the Risk Paradigm in Child Protection*, DOI:[10.1057/978-1-137-44130-0\\_3](https://doi.org/10.1057/978-1-137-44130-0_3).
20. Keddell E (2016) Substantiation decision-making and risk prediction in child protection systems. *Policy Quarterly*, DOI:[10.26686/pq.v12i2.4587](https://doi.org/10.26686/pq.v12i2.4587).
21. Keddell E (2026) Predictive risk decision-making tools in child welfare. *Elgar Encyclopedia of Social Work*, DOI:[10.4337/9781035310234.00032](https://doi.org/10.4337/9781035310234.00032).