IJAREEIE

e-ISSN: 2278-8875
p-ISSN: 2320-3765

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

Volume 15, Issue 2, February 2026

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.807

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering (IJAREEIE)

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.807 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 2, February 2026 ||

| DOI:10.15662/IJAREEIE.2026.1502009 |

Evaluating AI-Driven Risk Scoring Models for Child Protective Services Case Prioritization: Opportunities and Ethical Guardrails

Senthil Babu Malli Jayaraj

IT Project Manager, USA

ABSTRACT: Child protective services (CPS) agencies increasingly consider AI-driven risk scoring models to support intake screening and case prioritization decisions, informed by prominent implementations such as the Allegheny Family Screening Tool and subsequent tools adopted by other jurisdictions. These models offer the prospect of more consistent, evidence-informed prioritization at a time when intake call volume routinely exceeds caseworker capacity to fully investigate every referral with equal depth. At the same time, published research has raised substantial concerns regarding disparate impact, transparency, and due process when algorithmic scores influence decisions with such direct consequence for children and families.

This research article presents a governance framework for evaluating AI-driven risk scoring models in CPS case prioritization, examining model design and validation, disparate impact assessment, transparency and explainability requirements, human oversight design, and a five stage governance maturity model. The article draws on published algorithmic accountability research, child welfare practice literature, and prior responsible AI governance frameworks in this research series. Findings are illustrated with a distinct set of three-dimensional visualizations, including a 3D spiral scatter plot of risk score escalation across repeated referrals, a 3D binned density surface relating score to outcome, a 3D network graph of risk factor relationships, and a 3D stacked bar chart of subgroup composition by risk tier, alongside an extensive set of grid-style data tables.

Drawing on published child welfare algorithmic accountability research, NIST AI risk management guidance, and responsible AI governance literature through January 2026, this article finds that agencies applying structured disparate impact assessment, meaningful human oversight, and transparent case-level explanation report more defensible prioritization outcomes than agencies deploying scoring models without these governance practices, though published research also documents cases in which governance gaps have led to serious equity and due process concerns.

ETHICAL GUARDRAIL NOTE

This article is a governance and policy research analysis intended to help agencies evaluate and govern AI-driven risk scoring responsibly. It does not endorse any specific commercial or public model, and it does not provide technical guidance for building a risk scoring model. All statistics presented are illustrative unless directly attributed to a cited source.

KEYWORDS: child protective services, algorithmic risk scoring, predictive analytics, disparate impact, child welfare, responsible AI, due process, human oversight, case prioritization.

I. INTRODUCTION AND RESEARCH CONTEXT

Child protective services agencies process a high volume of referrals through intake hotlines, each requiring a screening decision about whether to open an investigation and, where an investigation opens, how urgently and thoroughly to pursue it. AI-driven risk scoring models, trained on historical administrative data to predict future substantiation, removal, or re-referral likelihood, have been piloted and, in some jurisdictions, deployed to support this screening and prioritization decision.

1.1 Research Objectives

- **Objective 1:** Characterize the operational case for AI-driven risk scoring in CPS intake and case prioritization.
- **Objective 2:** Examine model design, validation, and disparate impact assessment practice specific to the CPS context.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.807 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 2, February 2026 ||

| DOI:10.15662/IJAREEIE.2026.1502009 |

- **Objective 3:** Present transparency, human oversight, and due process requirements distinctive to child welfare decision-making.
- **Objective 4:** Provide a governance maturity model, role structure, and implementation roadmap for agencies evaluating or operating risk scoring tools.

.

1.2 Scope and Methodology

This research draws on the following source categories, all published prior to February 2026:

- Published algorithmic accountability research examining deployed child welfare risk scoring tools, including the Allegheny Family Screening Tool.
- Federal and state child welfare practice guidance addressing intake screening and case prioritization.
- NIST AI risk management framework guidance and responsible AI governance literature, extending prior responsible AI governance research in this series.
- Academic literature on algorithmic fairness, disparate impact measurement, and human-AI interaction in high-stakes public sector decisions.

Table 1 introduces the three illustrative risk scoring use points examined throughout this article.

Use Point	Description	Decision Informed
Intake screening	A risk score is generated at the point a referral is received, informing the screen-in or screen-out decision.	Whether to open an investigation at all.
Investigation prioritization	A risk score informs how quickly and with what resource intensity an accepted referral is investigated.	Sequencing and resourcing of accepted investigations.
Ongoing case risk reassessment	A risk score is periodically recalculated during an open case to inform continued service intensity.	Case planning and service intensity decisions for open cases.

1.3 Who This Article Is For

Audience	Primary Use of This Framework
Child welfare agency leadership	Deciding whether, and under what governance conditions, to pilot or deploy a risk scoring tool.
Data governance and analytics staff	Applying the validation and disparate impact assessment practices in Sections 5 and 6.
Caseworkers and supervisors	Understanding how a risk score should, and should not, influence professional judgment, per Section 9.
Legal, civil rights, and family advocacy stakeholders	Evaluating due process and transparency implications per Sections 8 and 10.

II. THE CASE FOR AI-DRIVEN RISK SCORING IN CPS INTAKE

Understanding the governance framework in this article requires first understanding, and taking seriously, the operational case that motivates agencies to consider AI-driven risk scoring in the first place.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.807 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 2, February 2026 ||

| DOI:10.15662/IJAREEIE.2026.1502009 |

2.1 Operational Drivers

Driver	Description	Illustrative Magnitude
Intake volume exceeding screening capacity	Hotline referral volume routinely exceeds the capacity for full, in-depth screening review of every referral.	Referral volumes in large jurisdictions frequently number in the tens of thousands annually.
Screening decision inconsistency	Published research has documented meaningful variation in screen-in and screen-out decisions across individual screeners reviewing similar referrals.	Documented variation has been substantial enough to motivate several jurisdictions' interest in structured decision support.
Desire for evidence-informed consistency	Structured tools, including but not limited to AI-driven scoring, aim to reduce unwarranted variation in how similarly situated referrals are handled.	Not unique to AI; also motivates non-AI structured decision-making tools long used in child welfare.

2.2 The Allegheny Family Screening Tool as Reference Context

The Allegheny Family Screening Tool, deployed by Allegheny County, Pennsylvania beginning in 2016, is the most extensively studied and publicly documented example of AI-driven risk scoring in CPS intake, and this article references it, and the substantial published research examining it, as context throughout rather than as an endorsement of any specific tool design.

Aspect	Description
Purpose	Generates a risk score at intake to support, not replace, the human screening decision.
Data inputs	Draws on integrated county administrative data spanning child welfare, criminal justice, and behavioral health systems, among others.
Publicly documented evaluation	Subject to multiple independent published evaluations examining both predictive performance and disparate impact, cited throughout Sections 5 and 6.
Governance posture	Deployed with an explicit human-in-the-loop design, in which the score informs but does not automatically determine the screening decision.

◇ PRACTICE NOTE

This article draws on the Allegheny Family Screening Tool as a widely studied reference point precisely because of its unusual degree of independent, published evaluation. Other jurisdictions' tools are generally less publicly documented, which itself is a governance consideration addressed in Section 8.

III. MODEL DESIGN AND INPUT FEATURE CONSIDERATIONS

The specific data inputs a risk scoring model draws on shape both its predictive performance and its disparate impact risk, making input feature selection a central governance decision rather than a purely technical one.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.807 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 2, February 2026 ||

| DOI:10.15662/IJAREEIE.2026.1502009 |

3.1 Common Input Feature Categories

Feature Category	Description	Governance Consideration
Prior CPS involvement history	Whether the family has prior referrals, investigations, or substantiated findings.	Strongly predictive in most published models, but risks perpetuating prior involvement patterns that themselves may reflect disparate historical screening or reporting.
Public system administrative data	Data drawn from integrated public systems such as behavioral health, criminal justice, or public benefits administration.	Raises equity concerns given documented disparities in surveillance intensity and system contact across populations, examined further in Section 6.
Household composition and demographic data	Information about household members' ages, relationships, and, in some designs, demographic characteristics.	Direct use of protected characteristics as model inputs is generally avoided or heavily scrutinized; even indirect proxies warrant disparate impact assessment.
Referral-specific narrative and allegation data	Details specific to the current referral, including allegation type and reporter-provided detail.	Generally less contested than administrative history data, though allegation categorization itself may vary systematically by source.

3.2 Feature Selection Principles

- **Principle 1:** Features should be selected based on demonstrated predictive validity for the specific outcome the model targets, not simply because the data happens to be available in an integrated data system.
- **Principle 2:** Features with a plausible causal or proxy relationship to protected characteristics warrant disparate impact assessment before inclusion, not only after a model is fully built, per the assessment practice in Section 6.
- **Principle 3:** Feature selection decisions and their rationale should be documented and retained for later audit, supporting the transparency requirements described in Section 8.

IV. RISK FACTOR RELATIONSHIPS AND COMPOSITE SCORING

Individual input features, per Section 3, combine into a composite risk score through a defined mathematical relationship, and understanding how features relate to one another and to the composite score is essential to governance review.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.807 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 2, February 2026 ||

| DOI:10.15662/IJAREEIE.2026.1502009 |

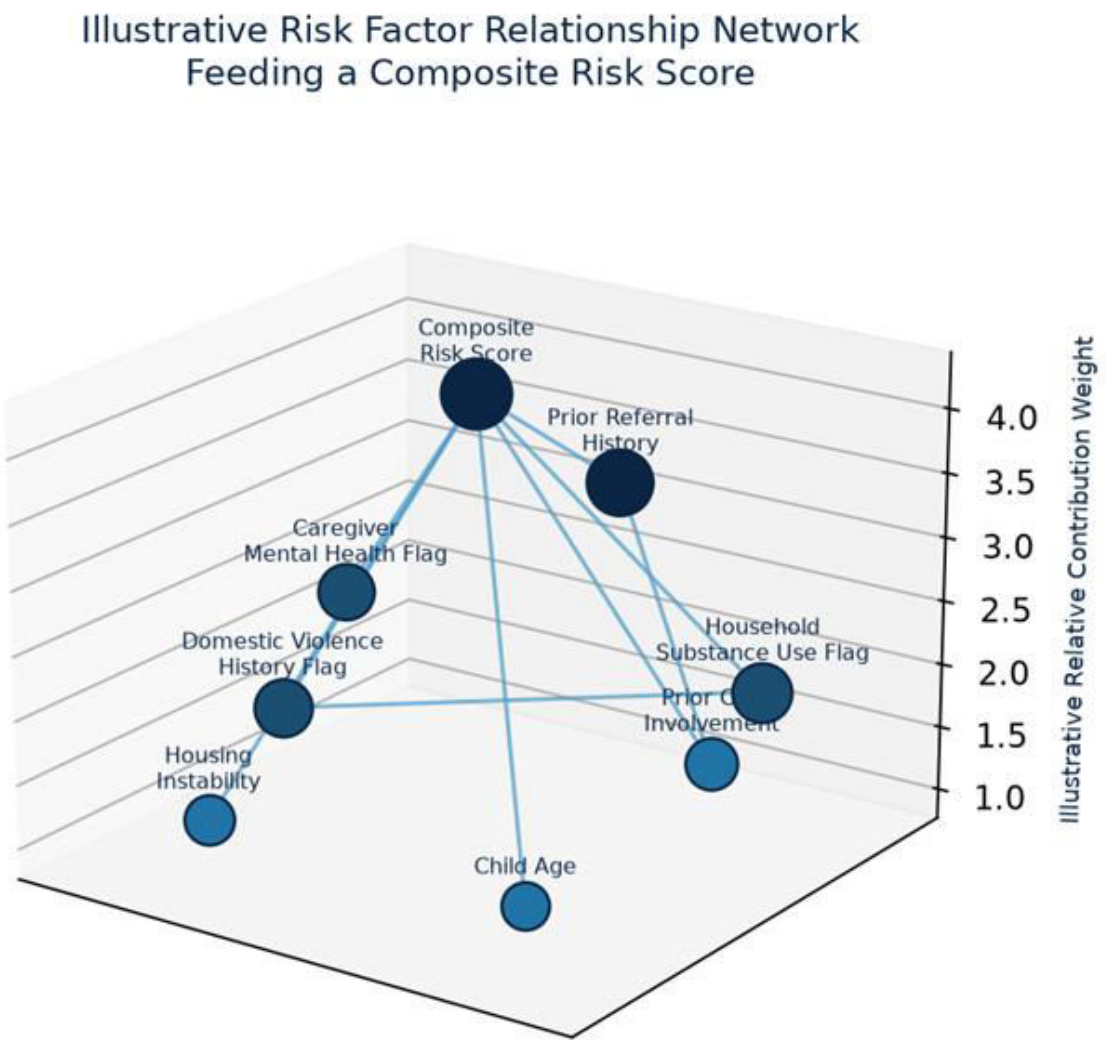


Figure.1. Illustrative risk factor relationship network feeding a composite risk score.

4.1 Interpreting the Risk Factor Network

As illustrated in Figure 1, several risk factors feed directly into the composite score, while some factors, such as prior referral history and prior CPS involvement, also relate to one another directly. This interconnection means that a single underlying condition, such as a family's extended prior system involvement, can influence the composite score through multiple correlated pathways simultaneously, potentially amplifying its effective weight beyond what a single feature's isolated contribution would suggest.

Risk Factor	Illustrative Relative Contribution Weight	Correlated With
Prior referral history	High	Prior CPS involvement
Household substance use flag	Moderate to high	Domestic violence history flag
Caregiver mental health flag	Moderate to high	No strong correlation illustrated
Domestic violence history flag	Moderate to high	Household substance use flag
Housing instability	Moderate	No strong correlation illustrated
Child age	Moderate	No strong correlation illustrated
Prior CPS involvement	Moderate to high	Prior referral history

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.807 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 2, February 2026 ||

| DOI:10.15662/IJAREEIE.2026.1502009 |

4.2 Governance Implications of Feature Correlation

- **Amplification risk:** Correlated features amplifying a single underlying condition's effective weight should be identified explicitly during model validation, per Section 5, rather than discovered only after deployment.
- **Compounding disparate impact risk:** Where correlated features both relate to documented disparate system contact patterns, examined in Section 6, their combined effect on disparate impact may exceed what either feature's individual disparate impact assessment would suggest.

V. MODEL VALIDATION AGAINST SUBSTANTIATED OUTCOMES

Model validation examines how well a risk score's predictions align with actual, later-confirmed outcomes, a foundational technical governance activity preceding any deployment decision.

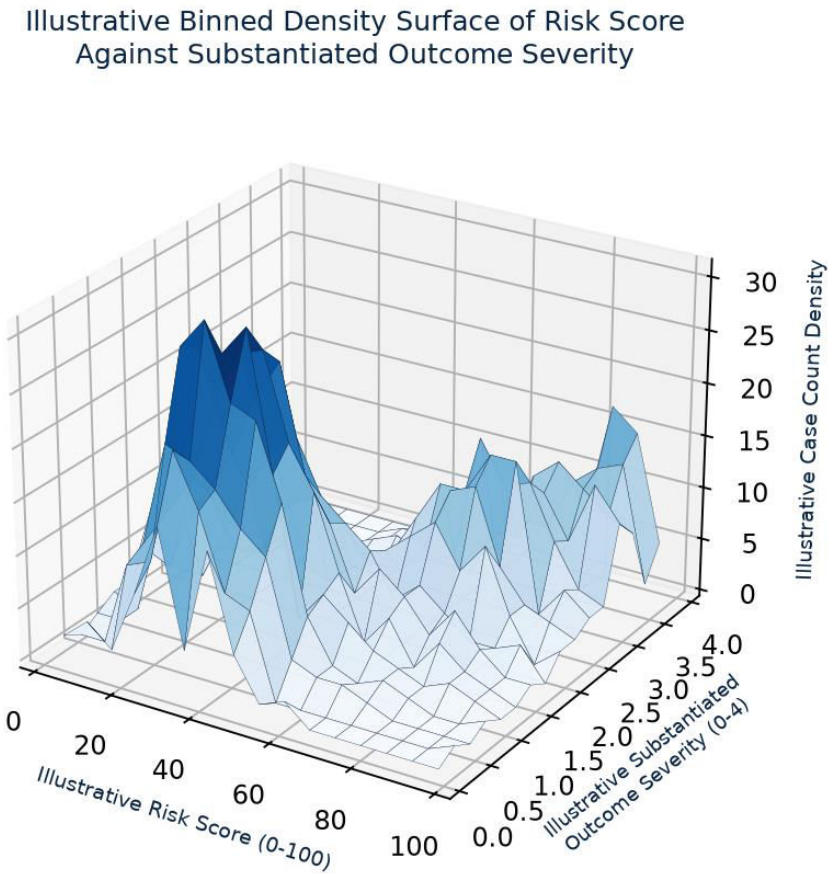


Figure.2. Illustrative binned density surface of risk score against substantiated outcome severity.

5.1 Interpreting the Density Surface

As illustrated in Figure 2, case density concentrates along a rising diagonal band, with higher risk scores generally associated with higher illustrated outcome severity, consistent with the score carrying genuine predictive signal. However, the band's width, rather than a narrow line, indicates substantial variation, meaning any individual case's actual outcome remains materially uncertain even at a given score level.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.807 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 2, February 2026 ||

| DOI:10.15662/IJAREEIE.2026.1502009 |

5.2 Validation Metric Comparison

Metric	Description	Interpretation Caution
Area under the receiver operating characteristic curve (AUC)	Measures the model's overall ability to rank higher-risk cases above lower-risk cases.	A high aggregate AUC does not guarantee acceptable performance within every demographic subgroup, addressed further in Section 6.
Positive predictive value at a given threshold	The percentage of cases flagged above a given score threshold that are later substantiated.	Should be evaluated at the actual threshold the agency intends to operationalize, not only at a generic reference threshold.
False negative rate	The percentage of later-substantiated cases that the model did not flag as high risk.	Carries particularly serious consequence in the CPS context given the safety stakes of a missed high-risk case.
Calibration	Whether a given score level corresponds consistently to the same actual outcome rate across the full case population.	Should be assessed by subgroup, not only in aggregate, per the disparate impact assessment in Section 6.

◇ PRACTICE NOTE
Published research evaluating deployed child welfare risk scoring tools has generally found meaningful, though imperfect, predictive validity, alongside important caveats regarding subgroup calibration and the appropriate scope of the specific outcome the model was validated to predict.

VI. DISPARATE IMPACT ASSESSMENT

Disparate impact assessment, examining whether a risk scoring model produces systematically different outcomes across demographic subgroups, is among the most consequential and most extensively debated governance activities in this domain, reflecting substantial published concern regarding racial and socioeconomic disparities in child welfare system contact generally.

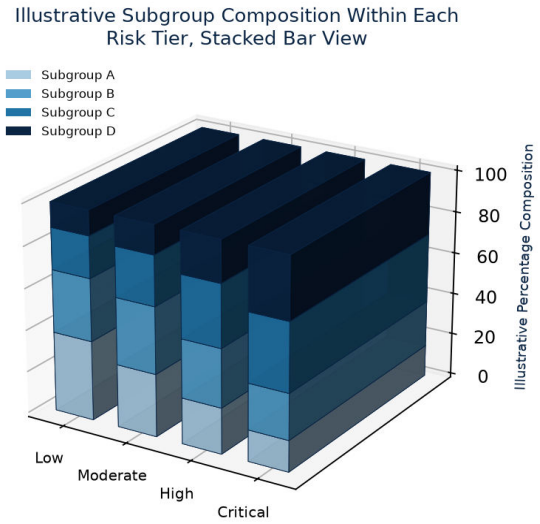


Figure.3. Illustrative subgroup composition within each risk tier, stacked bar view.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.807 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 2, February 2026 ||

| DOI:10.15662/IJAREEIE.2026.1502009 |

6.1 Interpreting the Subgroup Composition Pattern

As illustrated in Figure 3, subgroup composition shifts meaningfully across risk tiers, with Subgroup D's illustrated share rising from 12 percent in the Low tier to 30 percent in the Critical tier, while Subgroup A's illustrated share falls correspondingly from 38 percent to 15 percent. A shift of this kind, on its own, does not establish whether the pattern reflects genuine, defensible risk differences or reflects disparate historical system contact patterns embedded in the model's training data and input features.

Risk Tier	Subgroup A	Subgroup B	Subgroup C	Subgroup D
Low	38%	30%	20%	12%
Moderate	30%	32%	24%	14%
High	22%	28%	30%	20%
Critical	15%	22%	33%	30%

6.2 Disparate Impact Assessment Approaches

Approach	Description	Key Limitation
Subgroup calibration comparison	Comparing whether a given score level corresponds to the same actual substantiation rate across subgroups.	Requires sufficiently large subgroup sample sizes for statistically reliable comparison.
Disparate selection rate analysis	Comparing the rate at which different subgroups are flagged above a given action threshold.	A disparity in selection rate is not conclusive proof of bias if underlying, defensible risk genuinely differs, requiring careful, contextualized interpretation.
Root cause tracing to specific input features	Identifying which specific input features contribute most to an observed subgroup disparity.	Complex models can make individual feature contribution difficult to isolate cleanly, particularly where features are correlated per Section 4.2.
Comparison against pre-existing screening disparity	Comparing the model's subgroup disparity pattern against the disparity pattern of prior, non-AI-assisted screening decisions.	A model that merely reproduces pre-existing human decision disparity has not necessarily improved equity, even if it has not worsened it.

ETHICAL GUARDRAIL NOTE

Published research examining deployed child welfare risk scoring tools has reported mixed findings regarding disparate impact, with some studies finding a scoring tool's disparities comparable to or somewhat improved relative to pre-existing human screening decisions, and other research raising concerns about specific input features' disparate impact. Agencies should treat disparate impact assessment as an ongoing, rather than one-time, governance activity given this documented complexity, and should engage independent researchers or reviewers where feasible.

6.3 Structural Considerations Beyond the Model Itself

Disparate impact in CPS risk scoring cannot be fully addressed at the model level alone, since much of the underlying disparity reflected in administrative data originates upstream, in referral and reporting patterns the model did not create.

- **Upstream referral disparity:** Mandated reporters and the general public refer children to CPS hotlines at documented, differential rates across demographic and socioeconomic groups, a pattern that predates and is independent of any risk scoring model.
- **Inherited historical disparity:** A risk scoring model trained on historical substantiation data inherits any disparity embedded in how substantiation decisions were made historically, meaning model-level fairness interventions address only part of the overall equity picture.

VII. RISK SCORE ESCALATION ACROSS REPEATED REFERRALS

Families with multiple referrals over time present a distinctive pattern worth examining separately: how a risk score evolves as a family accumulates additional referral history.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.807 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 2, February 2026 ||

| DOI:10.15662/IJAREEIE.2026.1502009 |

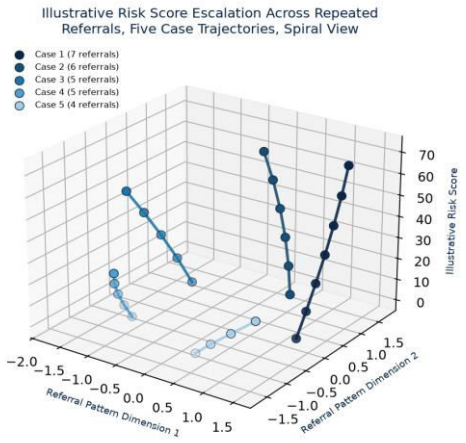


Figure.4. Illustrative risk score escalation across repeated referrals, five case trajectories, spiral view.

7.1 Interpreting the Escalation Pattern

As illustrated in Figure 4, each case trajectory rises across successive referrals, consistent with prior referral history contributing to the composite score per the risk factor network in Figure 1. The five illustrated trajectories rise at different rates, reflecting that referral recurrence alone does not uniformly predict risk; the specific nature and substantiation status of each prior referral matters as well.

- **Design consideration:** A model design that treats all prior referrals as equivalent, regardless of substantiation outcome, risks over-weighting families with high referral volume but low substantiation rates, a pattern sometimes associated with over-surveillance rather than genuine elevated risk.
- **Governance practice:** Escalation patterns should be reviewed periodically in aggregate to confirm that rising scores across repeated referrals correspond to genuinely rising risk rather than an artifact of how the model weights referral recurrence itself.

VIII. TRANSPARENCY AND EXPLAINABILITY REQUIREMENTS

Transparency and explainability, examined in general terms in prior responsible AI governance research in this series, carry particular weight in the CPS context given the direct, high-stakes consequences for families and the due process considerations examined in Section 10.

8.1 Transparency Requirements by Audience

Audience	Transparency Need	Current Practice Gap
Caseworkers using the score	Sufficient understanding of what drives a given score to exercise informed professional judgment about it, per Section 9.	Some deployed tools provide only the composite score without feature-level detail, limiting caseworker interpretability.
Affected families	Meaningful notice that a score informed a decision affecting them, and, where feasible, some understanding of the basis for that score.	Family-facing transparency is generally less developed than caseworker-facing transparency across published implementations.
Independent researchers and oversight bodies	Sufficient access to model design, validation, and disparate impact assessment data to conduct independent evaluation.	Varies substantially by jurisdiction; some tools, including the reference example in Section 2.2, have supported unusually extensive independent evaluation.
The general public	Basic understanding that risk scoring is in use and the general purpose it serves.	Public disclosure of risk scoring tool use varies; some jurisdictions provide detailed public documentation while others provide minimal disclosure.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.807 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 2, February 2026 ||

| DOI:10.15662/IJAREEIE.2026.1502009 |

8.2 Explainability Techniques Applicable to Risk Scoring

Technique	Description	Best Suited Context
Feature contribution reporting	Reporting which specific input features contributed most to a given case's score.	Caseworker-facing explanation supporting informed professional judgment per Section 9.
Plain-language risk category framing	Translating a numeric score into a plain-language risk category description avoiding false precision.	Family-facing and general public-facing communication.
Comparative case framing	Situating a given score relative to the overall population distribution, avoiding the impression of a definitive, individualized prediction.	Caseworker training and family-facing communication, tempering interpretation of any single score.

IX. HUMAN OVERSIGHT AND CASEWORKER DISCRETION

Human oversight design determines whether a risk score functions as one input informing professional judgment or, in practice, displaces that judgment, a distinction with direct consequence given the stakes involved.

9.1 Oversight Model Comparison

Oversight Model	Description	Risk
Score as one input among several	The score is presented alongside other case information, with the screener retaining full discretion to weigh it against other factors.	Requires screener training sufficient to genuinely integrate, rather than defer to, the score, echoing the automation bias risk examined in prior responsible AI research in this series.
Score as a mandatory floor or ceiling	The score sets a mandatory minimum or maximum response level that the screener cannot override without a defined exception process.	Reduces the risk of the score being ignored, but also reduces the scope for professional judgment to correct a score that is wrong for case-specific reasons.
Score as advisory only, prominently secondary	The score is available but deliberately de-emphasized in the screener's workflow relative to narrative case information.	May reduce automation bias risk but also may reduce the score's practical influence on decisions to the point where its intended consistency benefit is not realized.

9.2 Supporting Genuine, Not Nominal, Professional Judgment

- **Training design:** Screener training should explicitly address how and when professional judgment should override a risk score, supported by concrete examples, rather than leaving override criteria implicit or undefined.
- **Monitoring practice:** Override rates should be monitored over time; a persistently near-zero override rate may indicate automation bias, per the pattern examined in prior responsible AI governance research in this series, rather than a genuinely well-calibrated score requiring little correction.
- **Workload consideration:** Screener workload should be sufficient to allow genuine engagement with case-specific information beyond the score itself, since an overloaded screener facing high referral volume is more likely to default to the score regardless of its actual appropriateness for the specific case.

X. DUE PROCESS AND FAMILY NOTICE CONSIDERATIONS

Families affected by a risk-score-informed decision may have due process rights and interests distinct from, though related to, the transparency considerations examined in Section 8.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.807 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 2, February 2026 ||

| DOI:10.15662/IJAREEIE.2026.1502009 |

10.1 Due Process Considerations

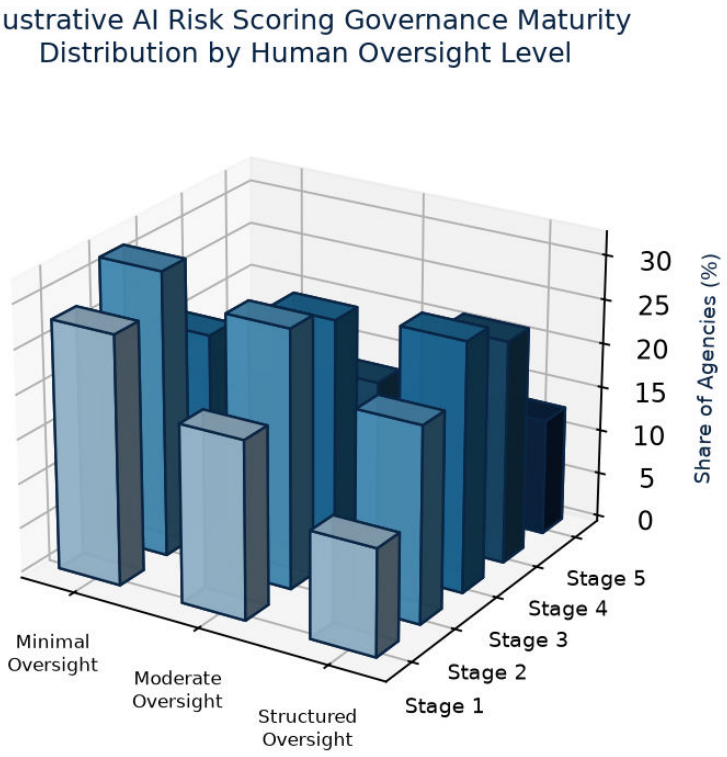
Consideration	Description	Governance Implication
Notice of algorithmic involvement	Whether affected families are informed that an algorithmic score played a role in a decision affecting them.	Practice varies significantly; agencies should establish and document a clear notice policy rather than leaving the question unaddressed.
Basis for challenge or correction	Whether a family has a meaningful avenue to challenge or seek correction of inaccurate underlying data feeding the score.	Data accuracy disputes should have a defined resolution process, particularly given that input features often draw on administrative data the family did not directly provide or verify.
Right to a human decision-maker	Whether a human retains genuine, not merely nominal, decision authority, connecting directly to the oversight design examined in Section 9.	Should be explicit in agency policy and reflected in actual practice, not only in policy documentation.

10.2 Legal and Policy Variation Across Jurisdictions

Due process requirements applicable to CPS decision-making are governed by federal constitutional doctrine, state statute, and state agency policy, and vary meaningfully across jurisdictions. This article's due process discussion is limited to identifying the categories of consideration relevant to risk scoring governance; agencies should engage legal counsel specific to their jurisdiction and program area when designing notice and challenge processes.

XI. THE FIVE STAGE GOVERNANCE MATURITY MODEL

Agencies can be classified into one of five maturity stages describing overall AI-driven risk scoring governance readiness, with maturity distribution varying by the degree of structured human oversight in place.



| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.807 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 2, February 2026 ||

| DOI:10.15662/IJAREEIE.2026.1502009 |

Figure.5. Illustrative AI risk scoring governance maturity distribution by human oversight level

Stage	Name	Description
1	Unvalidated	A scoring model, if used at all, has not undergone the structured validation described in Section 5.
2	Validated	Model validation per Section 5 has occurred, but disparate impact assessment per Section 6 is not yet formalized.
3	Assessed	Disparate impact assessment operates on a defined, recurring cadence, but human oversight design per Section 9 is not yet formalized.
4	Governed	Human oversight, transparency, and due process practices per Sections 8, 9, and 10 all operate under formal governance.
5	Continuously Reviewed	Ongoing monitoring per Section 13 and independent external review inform continuous governance refinement.

XII. GOVERNANCE ROLES AND OWNERSHIP

Sustained governance of AI-driven risk scoring in CPS requires roles spanning data science, child welfare practice, legal, and community stakeholder perspectives, given the multidimensional considerations examined throughout this article.

12.1 Governance Roles

Role	Primary Responsibility	Related Sections
Model validation and data science lead	Owns model validation methodology and ongoing performance monitoring.	Sections 5 and 13
Disparate impact assessment lead	Owns the recurring disparate impact assessment process and its documentation.	Section 6
Child welfare practice lead	Represents frontline case practice expertise in oversight design and screener training.	Section 9
Legal and due process reviewer	Reviews notice, challenge, and human decision-maker policy for legal sufficiency.	Section 10
Community and family advocacy liaison	Represents affected family and community perspective in governance review, a role distinct from internal agency staff.	Sections 6, 8, and 10

12.2 The Case for External and Community Perspective

Given the documented equity concerns examined in Section 6 and the direct consequence for families examined in Section 10, governance structures limited entirely to internal agency staff risk missing perspectives essential to fully evaluating a risk scoring tool's real-world impact.

- **Independent evaluation:** Independent academic or research partner evaluation, of the kind that has been conducted for the reference tool discussed in Section 2.2, provides a level of scrutiny internal review alone may not replicate.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.807 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 2, February 2026 ||

| DOI:10.15662/IJAREEIE.2026.1502009 |

- **Community representation:** Family and community advocacy representation in governance review, per Table 22, surfaces lived-experience perspective on transparency and due process practice that internal staff may not fully anticipate.

XIII. ONGOING MONITORING AND MODEL DRIFT

A risk scoring model's validated performance and assessed disparate impact at initial deployment do not guarantee sustained performance over time, as referral populations, reporting patterns, and underlying administrative data sources evolve.

13.1 Monitoring Practice Comparison

Monitoring Practice	Description	Recommended Cadence
Ongoing calibration monitoring	Confirming that score levels continue to correspond to the same actual outcome rates observed during initial validation, per Section 5.	Regular, with defined recalibration triggers.
Ongoing disparate impact monitoring	Repeating the disparate impact assessment described in Section 6 on a recurring basis rather than only at initial deployment.	At minimum annually, more frequently if triggered by monitoring indicators.
Override rate monitoring	Tracking screener override rates over time, per the practice described in Section 9.2.	Continuous, with periodic formal review.
Independent re-evaluation	Periodic evaluation by parties independent of the agency's own data science and program staff.	Periodically, informed by the agency's overall governance maturity stage per Section 11.

13.2 Common Drift Sources in the CPS Context

- **Referral pattern shifts:** Changes in mandated reporter training or public awareness campaigns can shift referral volume and composition in ways that affect model performance without any change to the model itself.
- **Upstream data source changes:** Changes to underlying administrative data systems feeding model input features, such as a change in how a partner agency codes a given data field, can silently affect model input quality.

XIV. RISK ASSESSMENT AND MITIGATION PLANNING

AI-driven risk scoring programs in CPS carry risks that agencies should evaluate and mitigate deliberately, given the profound consequences of both under- and over-intervention in family life.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.807 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 2, February 2026 ||

| DOI:10.15662/IJAREEIE.2026.1502009 |

Risk	Likelihood	Impact	Mitigation Approach
Disparate impact on specific subgroups goes undetected or unaddressed	Moderate to High	High	Establish recurring, independent disparate impact assessment per Sections 6 and 13.1.
Screeners defer to the score without genuine independent judgment (automation bias)	Moderate to High	High	Monitor override rates and design training per Section 9.2.
Insufficient family notice or challenge process	Moderate	High	Establish clear due process policy per Section 10.1, informed by legal review.
Model performance degrades due to upstream data or population drift	Moderate	Moderate to High	Establish ongoing monitoring per Section 13.1.
Governance limited to internal perspective misses real-world impact	Moderate to High	Moderate	Incorporate independent evaluation and community representation per Section 12.2.

14.1 Risk Prioritization Guidance

- Disparate impact and automation bias risks should be treated as highest priority given their direct connection to equity and to the fundamental purpose of maintaining genuine human professional judgment in high-stakes decisions.
- Due process and governance perspective risks are best addressed structurally, through the specific policy and representation practices described in Sections 10.1 and 12.2.

XV. IMPLEMENTATION ROADMAP

Agencies considering or operating AI-driven risk scoring should sequence validation, disparate impact assessment, oversight design, and governance formation deliberately, resisting pressure to deploy before foundational governance practices are in place.

Phase	Approximate Timeframe	Key Milestones
Model Design and Initial Validation	Months 1 to 6	Complete input feature selection per Section 3 and initial validation per Section 5.
Disparate Impact Assessment	Months 4 to 9	Complete initial disparate impact assessment per Section 6 before any production deployment decision.
Oversight and Transparency Design	Months 6 to 10	Design human oversight per Section 9 and transparency practices per Section 8, including screener training.
Due Process and Governance Formation	Months 8 to 12	Establish due process policy per Section 10 and formal governance roles per Section 12.
Pilot Deployment and Ongoing Monitoring	Months 12 onward	Deploy in a limited pilot context with the monitoring practices in Section 13 active from the outset.

15.1 Governance Cadence Reference

- **Monthly:** Review override rates and screener feedback per Section 9.2.
- **Quarterly:** Review calibration monitoring and any drift indicators per Section 13.1.
- **Annually:** Conduct formal disparate impact reassessment and a full governance maturity reassessment per Sections 6 and 11.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.807 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 2, February 2026 ||

| DOI:10.15662/IJAREEIE.2026.1502009 |

XVI. CASE PATTERNS AND PUBLISHED RESEARCH FINDINGS

This section synthesizes recurring patterns and findings from published research examining deployed or piloted child welfare risk scoring tools through early 2026, distinguishing established published findings from this article's own illustrative analysis elsewhere.

Pattern or Finding	General Direction of Published Evidence
Predictive validity of deployed tools relative to unstructured human screening alone	Multiple published evaluations report meaningful, though imperfect, predictive validity improvement.
Disparate impact relative to pre-existing human screening decisions	Findings are mixed; some evaluations report comparable or somewhat improved disparity, while other research raises specific concerns, particularly regarding features drawn from public system administrative data.
Screener trust and override behavior	Published research on human-AI interaction in high-stakes contexts generally documents a risk of over-reliance absent deliberate training and interface design, consistent with the automation bias concern in Section 9.2.
Transparency and public disclosure practice	Varies substantially by jurisdiction, with the reference tool discussed in Section 2.2 representing a comparatively well-documented example relative to typical practice.

16.1 Relationship to Adjacent Governance Research

The governance patterns examined in this article closely parallel structural lessons from responsible AI governance research examined elsewhere in this research series, applied here to the particularly high-stakes child welfare context.

Adjacent Governance Context	Parallel to This Article's Framework
Responsible AI adoption risk classification and governance gates	The five stage maturity model in Section 11 mirrors the risk-tiered governance gate structure examined in prior responsible AI governance research, adapted here to CPS-specific due process and disparate impact considerations.
AI-augmented vulnerability remediation human oversight design	The oversight model comparison in Section 9.1 mirrors the tiered human-in-the-loop design patterns examined in prior AI-augmented security tooling research, applied here to a substantially higher-stakes decision context.
Data quality and enterprise reporting architecture for CCWIS	The input feature governance discussion in Section 3 echoes the data quality dimension framework examined in prior CCWIS data governance research, extended here to address fairness in addition to accuracy.

XVII. KEY PERFORMANCE INDICATORS FOR ONGOING GOVERNANCE

Sustaining responsible AI-driven risk scoring governance requires an ongoing measurement framework spanning predictive validity, equity, and human oversight effectiveness.

KPI	Definition	Illustrative Target
Model calibration accuracy by subgroup	Whether score levels correspond to consistent actual outcome rates across subgroups, benchmarked against Section 6.	Consistent calibration across all subgroups; material divergence triggers review.
Screener override rate	Percentage of cases in which a screener's decision diverges from the score-suggested action, per Section 9.2.	Neither implausibly low nor unexpectedly high; consistent with genuine professional judgment.
Disparate impact assessment currency	Time since the most recent formal disparate impact assessment, per Section 13.1.	Within the annual or more frequent cadence described in Section 15.1.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.807 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 2, February 2026 ||

| DOI:10.15662/IJAREEIE.2026.1502009 |

Family notice and challenge process utilization	Rate at which affected families receive notice and, where applicable, exercise a challenge or correction process, per Section 10.1.	Consistent with the agency's documented due process policy.
Independent evaluation currency	Time since the most recent independent, external evaluation of the model, per Section 12.2.	Periodic, informed by governance maturity stage per Section 11.
Governance maturity stage progression	Composite score tracked against the five stage model in Section 11.	Steady progression toward higher maturity stages, with disparate impact and oversight practices prioritized early.

ETHICAL GUARDRAIL NOTE

KPIs should be reviewed by a governance body including the community and family advocacy representation described in Section 12.1, not solely by internal agency data science and program staff, given the equity and due process stakes examined throughout this article.

XVIII. CONCLUSION AND RECOMMENDATIONS

'AI-driven risk scoring offers CPS agencies a genuine opportunity to bring greater consistency to intake screening and case prioritization decisions made under significant volume pressure, but that opportunity carries commensurate governance responsibility given the profound stakes for children and families. Agencies that invest in rigorous validation, recurring disparate impact assessment, genuine human oversight, and meaningful transparency and due process practice are positioned to realize the operational benefits of risk scoring while managing its documented risks responsibly.

18.1 Summary of Key Findings

- Published evaluations of deployed child welfare risk scoring tools generally report meaningful predictive validity, but findings on disparate impact relative to pre-existing human screening are mixed and warrant continued, recurring assessment.
- Risk factor correlation can amplify the effective influence of a single underlying condition on a composite score, a pattern that model validation and disparate impact assessment should specifically examine.
- Human oversight design determines whether a risk score genuinely supports, or in practice displaces, professional judgment, with automation bias and screener workload as key influencing factors.
- Disparate impact in CPS risk scoring reflects both model-level design choices and upstream referral and reporting disparities that predate any specific model, meaning model-level intervention alone cannot fully resolve equity concerns.

18.2 Recommendations

- **Recommendation 1:** Complete rigorous model validation and disparate impact assessment before any production deployment decision, per Sections 5 and 6, rather than treating these as post-deployment activities.
- **Recommendation 2:** Design human oversight for genuine, monitored professional judgment, per Section 9, including explicit training on when and how to override a score.
- **Recommendation 3:** Establish clear, documented family notice and due process policy per Section 10, informed by jurisdiction-specific legal review.
- **Recommendation 4:** Incorporate independent evaluation and community or family advocacy representation into governance structures per Section 12.2, rather than relying solely on internal agency perspective.

REFERENCES

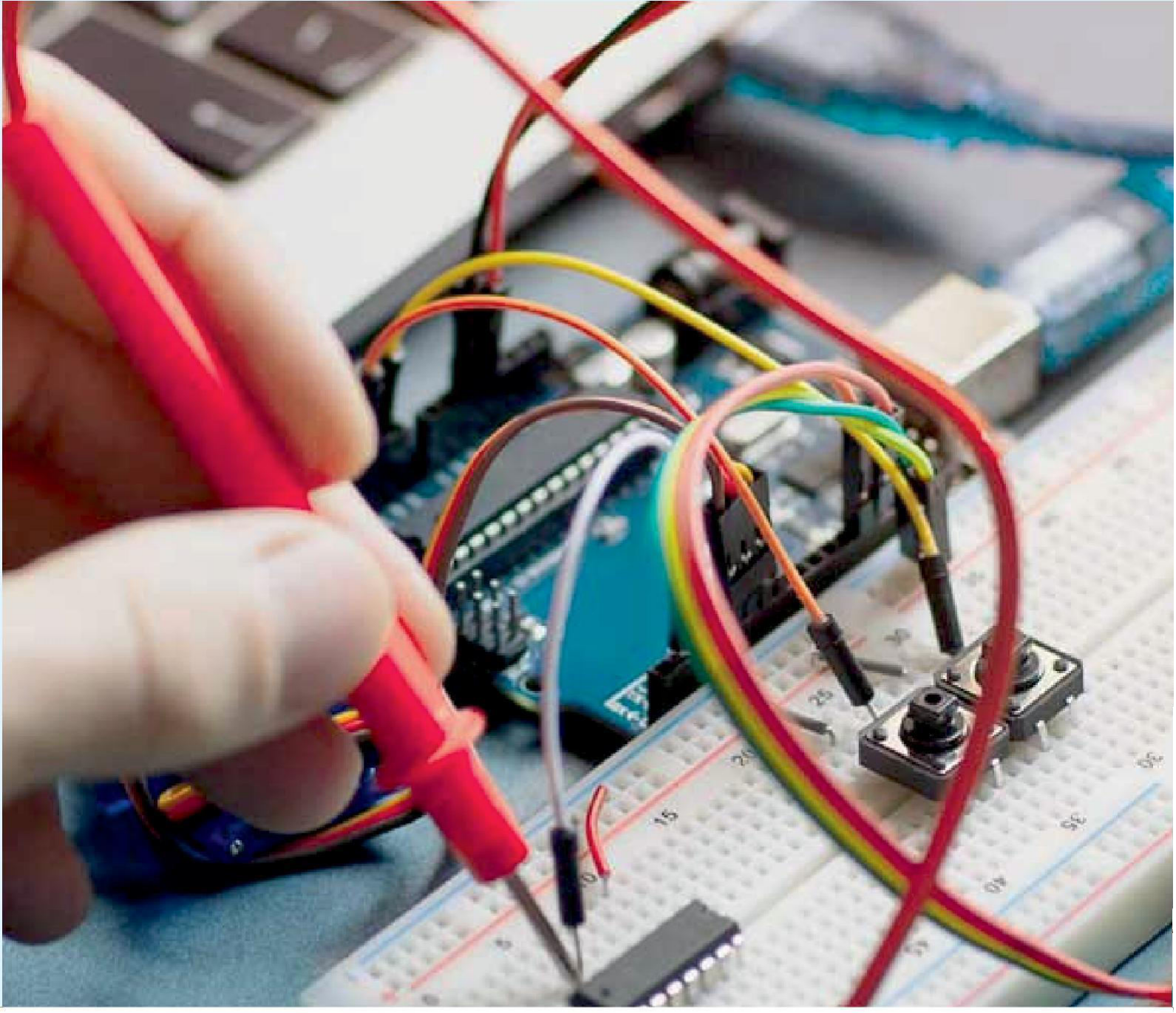
1. Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018). "A case study of algorithmic decision making in child maltreatment risk assessment." Conference on Fairness, Accountability, and Transparency (FAT*).
2. Vaithianathan, R., Maloney, T., Putnam-Hornstein, E., & Jiang, N. (2013). "Children in the public benefit system at risk of maltreatment: Identification via predictive modeling." American Journal of Preventive Medicine, 45(3), 354-359.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.807 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 2, February 2026 ||

| DOI:10.15662/IJAREEIE.2026.1502009 |

3. Hurley, D. (2018). "Can an Algorithm Tell When Kids Are in Danger?" The New York Times Magazine, January 2, 2018.
4. Eubanks, V. (2018). "Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor." St. Martin's Press.
5. Church, C.E., & Fairchild, A.J. (2017). "In Search of a Silver Bullet: Child Welfare's Embrace of Predictive Analytics." Juvenile and Family Court Journal, 68(1), 43-60.
6. American Civil Liberties Union (ACLU). (2021). "Family Surveillance by Algorithm: The Rapidly Spreading Tools Few Have Vetted." ACLU Report.
7. Allegheny County Department of Human Services. (2019). "Allegheny Family Screening Tool: Implementation and Evaluation." Allegheny County DHS Publication.
8. Shroff, R. (2017). "Predictive Analytics for Child Maltreatment Prevention." Data Science for Social Good, referenced in subsequent child welfare predictive analytics literature.
9. Gillingham, P. (2016). "Predictive Risk Modelling to Prevent Child Maltreatment and Other Adverse Outcomes for Service Users: Inside the 'Black Box' of Machine Learning." British Journal of Social Work, 46(4), 1044-1058.
10. Keddell, E. (2019). "Algorithmic Justice in Child Protection: Statistical Fairness, Social Justice and the Implications for Practice." Social Sciences, 8(10), 281.
11. National Institute of Standards and Technology (NIST). (2023). "Artificial Intelligence Risk Management Framework (AI RMF 1.0)." NIST Publication.
12. Barocas, S., Hardt, M., & Narayanan, A. (2019). "Fairness and Machine Learning: Limitations and Opportunities." fairmlbook.org.
13. Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). "Algorithmic Fairness: Choices, Assumptions, and Definitions." Annual Review of Statistics and Its Application, 8, 141-163.
14. U.S. Department of Health and Human Services, Administration for Children and Families. (2016). "Comprehensive Child Welfare Information Systems, Final Rule," 81 Fed. Reg. 35450, June 2, 2016.
15. Child Welfare Information Gateway. (2021). "Structured Decision Making in Child Welfare." U.S. Department of Health and Human Services, Children's Bureau.
16. U.S. Government Accountability Office. (2024). "Artificial Intelligence: Agencies Have Begun Implementation but Need to Complete Key Requirements." GAO Report.
17. Office of Management and Budget. (2024). "Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence," OMB Memorandum M-24-10.
18. Green, B., & Chen, Y. (2019). "Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments." Proceedings of the Conference on Fairness, Accountability, and Transparency.
19. Ho, D.E., & Xiang, A. (2020). "Affirmative Algorithms: The Legal Grounds for Fairness as Awareness." University of Chicago Law Review Online.
20. Casey Family Programs. (2018). "Data Quality and Governance in Child Welfare Information Systems." Casey Family Programs Research Report.



INNO SPACE
SJIF Scientific Journal Impact Factor

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

International Journal of Advanced Research

in Electrical, Electronics and Instrumentation Engineering

9940 572 462 6381 907 438 ijareeie@gmail.com

Scan to save the contact details