NCAC

CALiO™

CHILD ABUSE LIBRARY ONLINE

Using Artificial Intelligence in Child Protection and Child Welfare

A Bibliography

April 2026

Championing and Strengthening the Global Response to Child Abuse

nationalcac.org | 256-533-KIDS(5437) | 210 Pratt Avenue NE, Huntsville, AL 35801

Preferred citation: National Children's Advocacy Center. (2026). *Using artificial intelligence in child protection and child welfare: A bibliography.*

This project was supported by a grant awarded by the Office of Juvenile Justice and Delinquency Prevention, Office of Justice Programs, U.S. Department of Justice. Points of view or opinions in this document are those of the author and do not necessarily represent the official position or policies of the U.S. Department of Justice.

Scope

This bibliography covers literature examining issues related to the use of artificial intelligence (AI) and machine learning (ML) among professionals working in child protection and child welfare. While closely related, AI and ML are not interchangeable terms. AI refers to machines designed to perform tasks that typically require human intelligence. ML refers to systems that learn from data, recognizing patterns and improving their performance over time through experience. Issues covered in the literature include biases, reliability, trustworthiness, data privacy, AI literacy, ethics, and service outcomes.

Organization

Publications are listed in date descending order and include articles, book chapters, reports, research briefs, and international publications. Links are provided to full text publications when possible. However, this collection may not be complete. More information can be obtained in CALiO™, the Child Abuse Library Online.

Disclaimer

This bibliography was prepared by the Digital Information Librarians of the National Children's Advocacy Center (NCAC) for the purpose of research and education, and for the convenience of our readers. The NCAC is not responsible for the availability or content of cited resources. The NCAC does not endorse, warrant or guarantee the information, products, or services described or offered by the authors or organizations whose publications are cited in this bibliography. The NCAC does not warrant or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed in documents cited here. Points of view presented in cited resources are those of the authors, and do not necessarily coincide with those of the NCAC.

Using Artificial Intelligence in Child Protection & Child Welfare

A Bibliography

Gardiner, B., O'Donoghue, K., Yeung, P., & Jewel, Z. (2026). Social work practice and artificial intelligence: A scoping review. *Aotearoa New Zealand Social Work, 38*(1), 9-21. DOI:10.11157/anzswj-vol38iss1id1267

This scoping review explores artificial intelligence (AI) in social work between 2014 and 2024. It focuses on: a) the volume of the literature; b) the functions and benefits of AI in social work practice and the influence on social justice and social change on clients; c) the ethical considerations for social workers using AI; and d) the skills and knowledge social work practitioners need to have to navigate the AI-influenced landscape. Three electronic databases were searched, and 53 articles related to four research questions were identified. A PRISMA checklist and tool were followed. The articles were critically analysed for content, methodology, and findings through the research questions. The review found an emergent volume of literature indicating that AI applications in social work practice are heavily influenced by the AI model and design process implemented. Social workers' use of AI has both benefits and challenges. Ethical issues related to bias and social justice, as well as the reliability and trustworthiness of AI outcomes and decisions, were identified. The need for specific training in AI and social work to assist social workers in critically evaluating AI contributions and using them effectively was also noted. The article recommends further research to identify the prevalence of AI use in social work practice and its implications regarding ethics and social justice. It also recommends re-evaluating the scope of AI in social work and the need for social worker training.

Keddell, E., & Ballantyne, N. (2026). Social work & technology: Critical perspectives. *Aotearoa New Zealand Social Work, 38*(1), 1-8.

In our call for this special issue, we invited critical submissions exploring any aspect contemporary or historical—of the sometimes-troubled relationship between technology and social work. We were delighted by the range and quality of contributions received, and by the thoughtful ways in which authors have engaged with technology across education and practice. Unsurprisingly, many focus on artificial intelligence (AI), and generative AI (GenAI) in particular. Yet this is far from the only terrain covered in what is, ultimately, a wide-ranging and timely issue. We trust it will prove an invaluable resource for practitioners, educators, and researchers as we collectively orient ourselves to the latest wave of technological change—and reckon with its consequences for service users, social work agencies, and our fragile planet.

Van Katwyk, T., Banal, R., Seale, A., Bults, M., & Van Waterschoot, J. (2026). Critical participatory action for technology that can support transformative justice and change. *Aotearoa New Zealand Social Work, 38*(1), 75-86. DOI:10.11157/anzswj-vol38iss1id1285

As artificial intelligence, algorithms, and data-driven technologies become more embedded in social services, education, health care, and justice systems, concerns continue to grow about how these tools may reproduce existing social inequities. From a critical social work perspective, this article examines the ways in which artificial intelligence (AI) development and implementation can reinforce racism, ableism, classism, and other forms of structural oppression particularly through predictive and surveillance-based decision-making practices. The article explores how data-collection practices and algorithmic design are shaped by historical and institutional bias, despite frequent claims that these technologies are neutral or objective. Drawing on examples from criminal justice, child welfare, and health settings that utilised custom-built enterprise models, the analysis highlights the risks of opacity, universalisation, and

feedback loops that can deepen harm for marginalised individuals and communities. In response, the article advances an intersectional power analysis and proposes a critical participatory practice approach to technology development. By centring the knowledge, experiences, and leadership of those most impacted by AI-driven systems, this approach positions technology as a potential tool for transformational justice rather than social control. The article concludes by arguing that ethical engagement with AI in social work requires ongoing reflexivity, political accountability, and meaningful community participation.

Agbana, T. M. (2025). Ethical considerations in using predictive AI for risk assessment in child protection social work. *International Journal of Research Publication and Reviews, 6*(8), 1648-1663. DOI:10.55248/gengpi.6.0825.2922

Predictive artificial intelligence (AI) has the potential to help improve early intervention, resource allocation, and risk assessments in child protection social work. AI models can predict the likelihoods of harm happening based on case histories, demographic information and service interactions at large scales to allow social workers to intervene before an event occurs. However, this particular advance in technology also raises a host of highly complicated ethical questions that should be thoroughly explored, including questions of fairness, transparency, accountability and the delicate dance between public interest and individual rights on a broader level. This is particularly acute in the context of child protection, where errors can result in over-inclusiveness and unnecessary family interventions or under-inclusiveness that can have catastrophic consequences for children. Important ethical considerations are likely to include risks of algorithmic bias, perpetuating or worsening social inequalities in line with historical discrimination if the training data reflect systemic determinants; questions around data privacy and consent (child welfare is a sensitive domain); and the black-box algorithms undercutting trust and reducing professional discretion. Allowing algorithmic systems to relieve humans of the burden of risk assessment can also introduce novel forms of accountability and

questions about rights and due process when such decisions affect families. Overcoming these challenges, Picker, Metcalf, and Crawford argue will entail integrating ethical values in the design of AI systems, instituting meaningful mechanisms for accountability, engaging stakeholders, and enacting human-in the-loop solutions that allow for professional judgement. Through transparent governance frameworks governing predictive AI applications, child protection agencies can reap the benefits of technological tools without compromising the rights and dignity of the children and families they serve.

Ahn, E., Choi, M., Fowler, P., & Song, I. H. (2025). Artificial intelligence (AI) literacy for social work: Implications for core competencies. *Journal of the Society for Social Work and Research, 16*(1), 9-26. DOI:10.1086/735187

This paper discusses how efforts to build AI literacy can inform social work practice by exploring AI's impact on social inequalities across sectors and demonstrating how social work core competencies must evolve in response to AI-driven societal changes. Drawing on Long and Magerko's (2020) AI literacy framework, we propose integrating a social-work-specific AI literacy framework into the profession's core competencies to equip social workers to address emerging challenges while upholding professional values.

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, 169*, 107630. DOI:10.1016/j.chiabu.2025.107630

The objective was to understand the development and validation of contemporary machine learning (ML) models for child protection prediction, their performance evaluation, integration of fairness, and operationalisation of model explainability and transparency. This systematic review followed the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines. Model transparency was

assessed against the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis + Artificial Intelligence (TRIPOD+AI) criteria, while study risk of bias and model applicability were evaluated using Prediction model Risk of Bias ASsessment Tool (PROBAST) criteria. Eleven studies were identified, employing various ML approaches such as supervised classification models (e.g., binary classification, decision trees, support vector machines), regression models, and ensemble methods. These models utilised administrative health, child welfare, and criminal/court data. Performance was evaluated using a range of discrimination, classification, and calibration metrics, yielding variable results. Only 4 models incorporated group fairness, focusing on race/ethnicity as the protected attribute. Explainability and transparency were enhanced through Receiver Operating Curves, Precision-Recall Curves, feature importance plots, and SHapley Additive exPlanations (SHAP) plots. According to TRIPOD+AI criteria, only four studies reported likely reproducible models. Based on PROBAST criteria, all studies had unclear or high risk of bias. This is the first review to use TRIPOD+AI and PROBAST criteria to assess the risk of bias and transparency of ML models in child protection prediction. The findings reveal that the field remains methodologically immature, with many models lacking fair, transparent, and reproducible methods. Adoption of advanced fairness techniques, stakeholder involvement in model development and validation, and transparency through data and code sharing will be essential for the ethical and effective design of ML models, ultimately improving decision-making processes and outcomes for vulnerable children and families.

Chor, K. H. B., Luo, Z., Rodolfa, K. T., & Ghani, R. (2025). Exploring machine learning to support decision-making for placement stabilization and preservation in child welfare. *Journal of Child and Family Studies, 34*(1), 282-297. DOI:10.1007/s10826-024-02993-x

The Family First Prevention Services Act requires youth's placement in residential care to be clinically appropriate, time-limited, and only when youth's needs cannot be met in

family-like settings in foster care. State child welfare agencies can benefit from upstream, empirical decision support to preempt youth's placement disruption, coordinate proactive placement stabilization services, prevent unnecessary step-up to residential care, and improve outcomes for the youth. This statewide case study explores the potential benefit to child welfare decision support for placement stabilization and diversion from residential care, by comparing predictive machine learning (ML) models with conventional regression models. We analyzed child welfare spells of 12,621 youth in one large Midwestern state between January 2017 and January 2020. Caseworkers could refer youth to a placement stabilization and preservation program. To predict youth's monthly program need in the next 6 months, we developed and validated a wide grid of ML models—random forest, regularized logistic regression, decision tree, dummy classifier—and a conventional unregularized logistic regression model, using literature-informed predictors from child welfare administrative data. We retrained, retested, and compared all models over time using temporal hold-out sets. Based on anticipated program capacity, model evaluation focused on accuracy in identifying the 100 highest-need youth, fairness, and equity of resource allocation. Random forest models produced the best performance with a precision (positive predictive value) 10 times greater than baseline precision. Common important predictors across models included youth's age, history of placement changes, and emotional/behavioral needs. We discuss potential applications of ML to support preventive child welfare decisions, adapt to policy changes, and allocate limited resources.

Neeraj, M. S., & Nirmala, B. P. (2025). The role of digital libraries and emerging technologies in enhancing social work education: A comprehensive analysis. *Social Work Education, 44*(8), 1869-1884. DOI:10.1080/02615479.2024.2445740

Digital libraries refer to online repositories of academic resources, including books, journals, research papers, and multimedia, which provide unrestricted, virtual access to knowledge resources for educational and research purposes. The integration of digital

libraries and emerging technologies such as artificial intelligence (AI) and virtual reality (VR) is transforming social work education by enhancing access to information, fostering interdisciplinary learning, and promoting evidence-based practice. This paper explores the essential role of these tools in preparing social work students for the complexities of contemporary practice. While digital libraries expand access to critical resources, AI and VR offer personalized and immersive learning experiences that bridge the gap between theory and practice. The paper also addresses ethical considerations surrounding AI, such as issues of bias, privacy, and dependency, along with the need to contextualize these technologies globally. By examining the challenges and opportunities presented by digital libraries and emerging technologies, this analysis highlights their potential to shape a more inclusive, accessible, and effective approach to social work education.

Wilkins, D., & Benett, V. (2025). Making accurate judgements in child welfare: Comparing ChatGPT with qualified social workers. *Child & Family Social Work*. DOI:10.1111/cfs.13304

This study compares the judgemental accuracy of child and family social workers (n = 581) with ChatGPT, a generative AI model. Using 12 anonymized referrals, participants were asked predictive questions to evaluate accuracy through Brier scores. ChatGPT outperformed the average social worker on 11 of the 12 referrals, though the difference was not statistically significant. These findings highlight the potential and the limitations for AI to support decision-making in social work while emphasising the need to address ethical concerns and AI's inadequacies for understanding complex human needs and social contexts. The study contributes to ongoing discussions on integrating AI into social work, advocating for a balanced approach that enhances effectiveness while preserving the profession's essential human elements.

Yu, M. H., & Rose, R. A. (2025). Algorithmic-assisted decision-making tools in child welfare practice: A systematic review. *Research on Social Work Practice*. DOI:10.1177/10497315251350933

Algorithmic-assisted decision-making tools are increasingly used in child welfare services, yet key factors and challenges for successful and ethical implementation remain underexplored. This review centers on fairness, equity, and ethics in their application. Using PRISMA guidelines and including gray literature, nine studies were reviewed that described algorithmic tools implementation across various stages of the child welfare system. The review focused on identifying challenges and critical success factors, especially concerning fairness, equity, and ethics. This study used a holistic framework to review critical aspects from tool development to deployment. Additionally, strategies to address fairness and ethical considerations were identified and synthesized. Algorithmic-assisted tools hold promise for supporting high-stakes decisions in child welfare, but responsible use requires attention to ethical implementation. The review reveals significant methodological and empirical gaps, underscoring the need for future research to ensure equitable and effective deployment in practice.

Gibbs, D. J., Loper, A., Farley, A., Afkinich, J. L., Johnson, I. C., & Metz, A. J. (2024). Implementing algorithmic decision-making tools in child welfare systems: Practitioner perspectives on use and usefulness. *Journal of Technology in Human Services, 42*(4), 277-311. DOI:10.1080/15228835.2024.2402982

Decisions to screen child maltreatment reports are often inaccurate and inconsistent, which has prompted jurisdictions to develop algorithmic decision-making tools to supplement workers' judgments. However, the effectiveness of such innovations relies on successful adoption and consistent use by frontline users. Prior research has examined barriers to the adoption of decision-making tools in child welfare settings, but few studies have explored the implementation of algorithmic tools. This study described the use of such tools in practice and examined factors that influenced practitioners' attitudes and behaviors as they integrated the tools into their work. A qualitative case study informed

by the Technology Acceptance Model (TAM) and Consolidated Framework for Implementation Research (CFIR 2.0) was conducted regarding two county agencies implementing algorithmic tools for child welfare screening decisions. Data collection included document reviews, interviews with child welfare practitioners and leaders, and focus groups with child welfare and technology professionals. Participants disclosed key individual and contextual factors that impacted their perceptions of tool usefulness, including implementation processes, tool complexity, policy landscapes, internal communication structures, and staff role differences. Consideration of these factors must be incorporated into the future development and implementation of data-driven interventions to maximize their capacity to improve human services professionals' decision-making.

Ratner, H. F., & Schrøder, I. (2024). Ethical plateaus in Danish child protection services: The rise and demise of algorithmic models. *Science and Technology Studies, 37*(3), 44-61. DOI:10.23987/sts.126011

This paper analyses how controversies shape an emerging field of AI in Danish child protection services. In a context of high controversiality, we examine how algorithmic systems evolve in conjunction with changing ethical stakes. Empirically, we report a study comprising all Danish attempts (n=4) to develop algorithmic models for child protection services. These attempts were never fully implemented and have been either cancelled, paused or changed significantly since their outset. Combining the notion of 'ethical plateaus' with insights from valuation studies, we propose that public controversies shape how organisations enact their algorithms as ethically 'good'. Our findings demonstrate how valuations of ethically contestable algorithms involve the very distribution of agency across humans and algorithms, i.e., how much power and agency should be delegated to algorithmic models. In the case of Danish child protection services, this moves towards reducing their agency.

Saxena, D., & Guha, S. (2024). Algorithmic harms in child welfare: Uncertainties in practice, organization, and street-level decision-making. *ACM Journal on Responsible Computing, 1*(1), 1-32. DOI:10.1145/3616473

Algorithms in public services such as child welfare, criminal justice, and education are increasingly being used to make high-stakes decisions about human lives. Drawing upon findings from a two-year ethnography conducted at a child welfare agency, we highlight how algorithmic systems are embedded within a complex decision-making ecosystem at critical points of the child welfare process. Caseworkers interact with algorithms in their daily lives where they must collect information about families and feed it to algorithms to make critical decisions. We show how the interplay between systemic mechanics and algorithmic decision making can adversely impact the fairness of the decision-making process itself. We show how functionality issues in algorithmic systems can lead to process-oriented harms where they adversely affect the nature of professional practice, and administration at the agency, and lead to inconsistent and unreliable decisions at the street level. In addition, caseworkers are compelled to undertake additional labor in the form of repair work to restore disrupted administrative processes and decision-making, all while facing organizational pressures and time and resource constraints. Finally, we share the case study of a simple algorithmic tool that centers caseworkers' decision-making within a trauma-informed framework and leads to better outcomes, however, required a significant amount of investments on the agency's part in creating the ecosystem for its proper use.

Schlögl-Flierl, K., & Ziethmann, P. (2024). Addressing the needs and demands of child welfare: A connection between AI ethics and ethics of vulnerability. In M. Reder & C. Koska (Eds.), *Künstliche intelligenz und ethische verantwortung* (pp. 85-100). Bielefeld. DOI:10.14361/9783839469057-006

The aim of this paper is to bring a critical position into the debate about the use of AI in child welfare. However, the focus should not be on a false faith in technology. (cf.Klöcker2021) The human judgement is not unchallenged, so we are not arguing here

that decisions should exclusively be made by humans.(cf. Schwabe 2022) For example, a study supported by the Federal Anti-Discrimination Agency has stressed, with regard to learning algorithms in particular, that in many cases human decisions can be the sources of discrimination risks (cf. Orwat 2019).

Pesonen, K., Korpela, J., Vilko, J., & Elfvengren, K. (2023). Realizing the value potential of AI in service needs assessment: Cases in child welfare and mental health services. *Proceedings of the 56th Hawaii International Conference on System Sciences,* 2840-2849. https://scholarspace.manoa.hawaii.edu/communities/f36bf371-28c7-4427-bff7-718d2c995872

In social and health care the use of technology that utilizes data has great potential from the point of view of value creation. This case study examines the factors that impact the value potential realization of AI prediction models as part of the customer/patient service need assessment process. The research focuses on a pilot project of a Finnish case organization, in which prediction models were tested in child welfare and mental health services. Both positive and negative value-realizing factors were found in the research. The information produced by artificial intelligence has great value potential. Regulation and transparency of data need to be addressed, but at the same time, more flexible use of social and health register data needs to be considered to ensure that resources are allocated in a value-added way.

Reamer, F. G. (2023). Artificial intelligence in social work: Emerging ethical issues. *International Journal of Social Work Values and Ethics, 20*(2), 52-71. DOI:10.55521/10-020-200

Artificial intelligence (AI) is becoming increasingly prevalent in social work. AI is being used to conduct risk assessments, assist people in crisis, strengthen prevention efforts, identify systemic biases in the delivery of social services, provide social work education, and predict social worker burnout and service outcomes, among other uses. There is now

considerable literature on the ways in which social workers and other human service professionals can use AI to assist vulnerable people. Yet social work's literature does not include in-depth examination of the ethical implications of practitioners' use of AI. The purpose of this article is to examine ethical issues related to social workers' use of AI; apply relevant ethical standards; and outline elements of a strategy for social workers' ethical use of AI. Key ethical issues addressed include informed consent and client autonomy; privacy and confidentiality; transparency; client misdiagnosis; client abandonment; client surveillance; plagiarism, dishonesty, fraud, and misrepresentation; algorithmic bias and unfairness; and use of evidence-based AI tools.

Trudeau, K. J., Yang, J., Di, J., Lu, Y., & Kraus, D. R. (2023). Predicting successful placements for youth in child welfare with machine learning. *Children and Youth Services Review, 153*, 107117. DOI:10.1016/j.childyouth.2023.107117

Out-of-home placement decisions have extremely high stakes for the present and future well-being of children in care because some placement types, and multiple placements, are associated with poor outcomes. We propose that a clinical decision support system (CDSS) using existing data about children and their previous placement success could inform future placement decision-making for their peers. The objective of this study was to test the feasibility of developing machine learning models to predict the best level of care placement (i.e., the placement with the highest likelihood of doing well in treatment) based on each youth's behavioral health needs and characteristics. We developed machine learning models to predict the probability of each youth's treatment success in psychiatric residential care (i.e., Psychiatric Residential Treatment Facility [PRTF]) versus any other placement (AUROCs > 0.70) using data collected in standard care at a behavioral health organization. Placement recommendations based on these machine learning models distinguished between youth who did well in residential care versus non-residential care (e.g., 80% of those who received care in the recommended setting with the highest predicted likelihood of success had above average risk-adjusted outcomes).

Then we developed and validated machine learning models to predict the probability of each youth's treatment success across specific placement types in a state-wide system, achieving an average AUROC score of >0.75. Machine learning models based on risk-adjusted behavioral health and functional data show promise in predicting positive placement outcomes and informing future placement decisions for youth in care. Related ethical considerations are discussed.

Blanchard, M. (2022). Predictive analytics in child welfare: Five principles for regulating algorithmic accountability in a new wave of predictive models. *University of Baltimore Law Review, 51*(3), 5.

This comment discusses the issue of child maltreatment and gives a historical overview of the various methods used for screening abuse and neglect. It outlines how calls are screened by human case workers and how deficiencies in that system led some jurisdictions to introduce predictive analytics into the process. Part III provides an analysis of the new Hello Baby predictive risk modeling (PRM) tool and discusses how, despite improved efforts, there are still many deficiencies and potential discrimination within the model. This comment explores various principles the Hello Baby model could adopt to help improve accuracy and limit bias. This comment concludes by summarizing the next steps that must be taken if Allegheny County or other jurisdictions hope to see the Hello Baby PRM tool and similar models work successfully and with limited disproportionate effects on minority groups and low-income families.

Jørgensen, A. M., & Nissen, M. A. (2022). Making sense of decision support systems: Rationales, translations and potentials for critical reflections on the reality of child protection. *Big Data & Society, 9*(2), 1-13. DOI:10.1177/20539517221125163

Decision support systems, which incorporate artificial intelligence and big data, are receiving significant attention in the public sector. Decision support systems are

sociocultural artefacts that are subject to a mix of technical and political choices, and critical investigation of these choices and the rationales they reflect are paramount since they are inscribed into and may cause harm, violate fundamental rights and reproduce negative social patterns. Applying and merging the concepts of sense-making and translation, this article investigates the rationales, translations and critical reflections that shape the development of a decision support system to support social workers assessing referrals concerning child neglect. It presents findings from a qualitative case study conducted in 2019–2020 at the Citizen Centre Children and Young People, Copenhagen Municipality, Denmark. The analysis shows how key actors through processes of translation construct, negotiate and readjust problem definitions, roles, interests, responsibilities and ideas of ambiguity and accountability. Although technological solutionism is present in these processes, it is not the only rationale invested. Rather, technological and data-driven rationales are adjusted to and merged with rationales of efficiency, return on investment and child welfare. Through continuous renegotiation of roles, responsibilities and problems according to these rationales, the key actors attempt to orchestrate ways of managing the complexity facing child welfare services by projecting images of future potentials of the decision support system that are yet to be realised.

Kawakami, A., Sivaraman, V., Cheng, H. F., Stapleton, L., Cheng, Y., Qing, D., Perer, A., Wu, Z. S., Zhu, H., & Holstein, K. (2022). Improving human-AI partnerships in child welfare: Understanding worker practices, challenges, and desires for algorithmic decision support. In S. Barbosa, C. Lampe, C. Appert, D. A. Shamma, S. Drucker, J. Williamson, & K. Yatani (Eds.), *Proceedings of the 2022 CHI conference on human factors in computing systems* (pp. 1-18). Association for Computing Machinery. DOI:10.1145/3491102.3517439

AI-based decision support tools (ADS) are increasingly used to augment human decision-making in high-stakes, social contexts. As public sector agencies begin to adopt ADS, it is critical that we understand workers' experiences with these systems in practice.

In this paper, we present findings from a series of interviews and contextual inquiries at a child welfare agency, to understand how they currently make AI-assisted child maltreatment screening decisions. Overall, we observe how workers' reliance upon the ADS is guided by (1) their knowledge of rich, contextual information beyond what the AI model captures, (2) their beliefs about the ADS's capabilities and limitations relative to their own, (3) organizational pressures and incentives around the use of the ADS, and (4) awareness of misalignments between algorithmic predictions and their own decision-making objectives. Drawing upon these findings, we discuss design implications towards supporting more effective human-AI decision-making.

Stapleton, L., Lee, M. H., Qing, D., Wright, M., Chouldechova, A., Holstein, K., Wu, Z. S., & Zhu, H. (2022). Imagining new futures beyond predictive systems in child welfare: A qualitative study with impacted stakeholders. *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency* (pp. 1162-1177). Association for Computing Machinery. DOI:10.1145/3531146.3533177

Child welfare agencies across the United States are turning to datadriven predictive technologies (commonly called predictive analytics) which use government administrative data to assist workers' decision-making. While some prior work has explored impacted stakeholders' concerns with current uses of data-driven predictive risk models (PRMs), less work has asked stakeholders whether such tools ought to be used in the first place. In this work, we conducted a set of seven design workshops with 35 stakeholders who have been impacted by the child welfare system or who work in it to understand their beliefs and concerns around PRMs, and to engage them in imagining new uses of data and technologies in the child welfare system. We found that participants worried current PRMs perpetuate or exacerbate existing problems in child welfare. Participants suggested new ways to use data and data-driven tools to better support impacted communities and suggested paths to mitigate possible harms of these tools. Participants also suggested low-tech or no-tech alternatives to PRMs to address problems in child welfare. Our study sheds light on how researchers and designers can

work in solidarity with impacted communities, possibly to circumvent or oppose child welfare agencies.

Goldkind, L. (2021). Social work and artificial intelligence: Into the matrix. *Social Work, 66*(4), 372-374. DOI:10.1093/sw/swab028

The advent of these highly automated tools has prompted a demand from academic, industry, and government sectors to examine how digital decision making can act to concentrate human bias. This call to infuse ethics and social justice–centered design into computer and data science curricula and the technology sector represents a significant opportunity for social work. As a values-centered profession with a robust code of ethics, social work is uniquely positioned to engage across disciplines to inform the creation of thoughtful algorithmically enhanced policy and practice at all levels. Social work's core values of social justice, integrity, and the primacy of relationships render us uniquely suited to assist developers as they empirically test the effectiveness of their algorithmic products. Our ethical duty to vulnerable populations requires that we monitor and assess the data and the assumptions used to train these algorithms, attending to the social implications of an emerging generation of tools.

Lappalainen, K. F. (2021). Protecting children from maltreatment with the help of artificial intelligence: A promise or a threat to children's rights?. In K. de Vries & M. Dahlberg (Eds.), *Law, AI, and digitalization* (pp. 431-466). De Lege.

Contrary to the promise and hopes for AI tools is the fact that the use of such tools for child protection, comes with multiple risks from a children's rights perspective. This is certainly the case regarding the use of predictive risk modelling (PRM) in child welfare. This chapter examines the benefits and risks of using AI tools for child protection.

Vaithianathan, R., Benavides-Prado, D., Dalton, E., Chouldechova, A., & Putnam-Hornstein, E. (2021). Using a machine learning tool to support high-stakes decisions in child protection. *AI Magazine, 42*(1), 53-60.

Machine learning decision support tools have become popular in a range of criminal justice, and child welfare. But the design of these tools often fails to consider the potentially complex interactions that happen between the tools and humans. This lack of human centered design is one reason that so few tools are actually deployed, and even if they are, struggle to achieve impact. In this article we present the example of the Allegheny Family Screening Tool, a machine learning model used since 2016 to support hotline screening of child maltreatment referrals. We describe aspects of human-centered design that contributed to the successful deployment of this tool, including agency leadership and ownership, transparency by design, ethical oversight, community engagement, and social license. Finally, we identify potential next-steps to encourage greater integration of human-centered design into the development and implementation of machine learning decision support tools.

Victor, B. G., Perron, B. E., Sokol, R. L., Fedina, L., & Ryan, J. P. (2021). Automated identification of domestic violence in written child welfare records: Leveraging text mining and machine learning to enhance social work research and evaluation. *Journal of the Society for Social Work and Research, 12*(4), 631-655. DOI:10.1086/712734

Child welfare agencies often lack information about the front-end service needs of the families they serve. Thus, the current study tests the feasibility of text mining and machine learning procedures for identifying problems related to domestic violence documented in child welfare investigation summaries. Method: We labeled child welfare investigation summaries (N = 1,402) for the presence or absence of an active domestic violence service need. Labeled documents were then used to develop text mining and machine learning models and test their accuracy and reliability. Results: Machine learning models achieved greater than 90% accuracy when compared with human coders. Fleiss kappa estimates of coding reliability between the top-performing model and human reviewers exceeded

.80, indicating that our model could support human reviewers to complete this coding task. Conclusion: Results provide strong evidence that text mining and machine learning procedures can be a cost-effective solution for extracting meaningful insights from text data. Although unsuitable for case-level predictive analytics, insights derived from these procedures can be particularly useful for investigating the prevalence, temporal trends, and geographic distribution of domestic violence-related needs in the child welfare system. These methods could substantially enhance the use of text data in social work research and evaluation.

Henman, P. (2020). Improving public services using artificial intelligence: Possibilities, pitfalls, governance. *Asia Pacific Journal of Public Administration, 42*(4), 209-221. DOI:10.1080/23276665.2020.1816188

Artificial intelligence arising from the use of machine learning is rapidly being developed and deployed by governments to enhance operations, public services, and compliance and security activities. This article reviews how artificial intelligence is being used in public sector for automated decision making, for chatbots to provide information and advice, and for public safety and security. It then outlines four public administration challenges to deploying artificial intelligence in public administration: accuracy, bias and discrimination; legality, due process and administrative justice; responsibility, accountability, transparency and explainability; and power, compliance and control. The article outlines technological and governance innovations that are being developed to address these challenges.

Saxena, D., Badillo-Urquiola, K., Wisniewski, P. J., & Guha, S. (2020). A human centered review of algorithms used within the US child welfare system. *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1-15). Association for Computing Machinery. DOI:10.1145/3313831.3376229

The U.S. Child Welfare System (CWS) is charged with improving outcomes for foster youth; yet, they are overburdened and underfunded. To overcome this limitation, several states have turned towards algorithmic decision-making systems to reduce costs and determine better processes for improving CWS outcomes. Using a human-centered algorithmic design approach, we synthesize 50 peer-reviewed publications on computational systems used in CWS to assess how they were being developed, common characteristics of predictors used, as well as the target outcomes. We found that most of the literature has focused on risk assessment models but does not consider theoretical approaches (e.g., child-foster parent matching) nor the perspectives of caseworkers (e.g., case notes). Therefore, future algorithms should strive to be context-aware and theoretically robust by incorporating salient factors identified by past research. We provide the HCI community with research avenues for developing human-centered algorithms that redirect attention towards more equitable outcomes for CWS.

Schafer, J. G. (2020). Harnessing AI innovation for struggling families. *University of Illinois Journal of Law, Technology, and Policy, 2020*(1), 411-455.

This article will (1) provide an overview of the data that is currently stored and collected by state child welfare systems; (2) describe the complex set of *412 federal laws that restrict sharing of this information and propose legislative and regulation changes that, if implemented, would foster technological innovation that could be life-changing for children and families; (3) suggest currently feasible machine learning applications that would benefit tech-optimized child welfare systems; (4) describe the practical steps needed to ready state child welfare agencies to implement technology innovation; and (5) analyze privacy considerations in adopting AI technologies.

Keddell, E. (2019). Algorithmic justice in child protection: Statistical fairness, social justice and the implications for practice. *Social Sciences, 8*(10), 281. DOI:10.3390/socsci8100281

Algorithmic tools are increasingly used in child protection decision-making. Fairness considerations of algorithmic tools usually focus on statistical fairness, but there are broader justice implications relating to the data used to construct source databases, and how algorithms are incorporated into complex sociotechnical decision-making contexts. This article explores how data that inform child protection algorithms are produced and relates this production to both traditional notions of statistical fairness and broader justice concepts. Predictive tools have a number of challenging problems in the child protection context, as the data that predictive tools draw on do not represent child abuse incidence across the population and child abuse itself is difficult to define, making key decisions that become data variable and subjective. Algorithms using these data have distorted feedback loops and can contain inequalities and biases. The challenge to justice concepts is that individual and group rights to non-discrimination become threatened as the algorithm itself becomes skewed, leading to inaccurate risk predictions drawing on spurious correlations. The right to be treated as an individual is threatened when statistical risk is based on a group categorisation, and the rights of families to understand and participate in the decisions made about them is difficult when they have not consented to data linkage, and the function of the algorithm is obscured by its complexity. The use of uninterpretable algorithmic tools may create 'moral crumple zones', where practitioners are held responsible for decisions even when they are partially determined by an algorithm. Many of these criticisms can also be levelled at human decision makers in the child protection system, but the reification of these processes within algorithms render their articulation even more difficult, and can diminish other important relational and ethical aims of social work practice.

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

This paper presents the findings from a study designed to explore whether predictive analytics and machine learning could improve the accuracy and utility of the child welfare risk assessment instrument used in Broward County (Ft. Lauderdale, Florida). The findings from this study indicate that, indeed, predictive analytics and machine learning would significantly improve the accuracy and utility of the child welfare risk assessment instrument being used. If the predictive analytic and machine learning algorithms developed in this study would be deployed, there would be improved accuracy in identifying low, moderate and high-risk cases, better matching between the needs of children and families and available services and improved child and family outcomes. This paper also identifies further areas of research and study.