Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An Ethical-Legal Framework for AI Early Warning Systems in Trafficking Prevention

Priscilla Kuyet Ishaku

ishakupriscillak@gmail.com

Birmingham City University

Henrietta Chukwufumnanya Ikediashi

Jaipur National University

Elizabeth Ikorrishor Igbodor

Northumbria University

Research Article

Keywords: Artificial intelligence ethics, human trafficking, forced marriage, predictive surveillance, algorithmic accountability

Posted Date: June 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9801239/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

The deployment of artificial intelligence for preventing human trafficking and forced marriage offers transformative potential yet raises profound ethical and legal challenges that demand rigorous interdisciplinary scrutiny. Current interventions remain predominantly reactive, leaving a critical gap in preventive governance that respects international human rights obligations. This paper proposes an integrated ethical-legal framework for AI-powered early warning systems designed to identify at-risk individuals before harm occurs. Drawing on a systematic interdisciplinary review, doctrinal legal analysis, and comparative ethical assessment, we examine predictive analytics, natural language processing, and geospatial monitoring tools through the lenses of feminist legal theory, human security, and the EU AI Act. Our analysis reveals significant risks of algorithmic bias, surveillance overreach, and accountability gaps that disproportionately affect marginalised women and children. We present a six-component governance framework that embeds human-in-the-loop oversight, mandatory bias auditing, privacy-by-design, and child-centred safeguards. The findings demonstrate that legitimacy in AI-enhanced prevention depends not only on technical efficacy but also on embedding ethical safeguards and legal compliance throughout the algorithmic lifecycle. This research contributes to the technology-law-society discourse by bridging AI ethics with international human rights standards and offering actionable guidance for policymakers, developers, and practitioners.

1. Introduction

Human trafficking and forced marriage persist as among the most severe violations of women's and children's rights in the twenty-first century. The United Nations Office on Drugs and Crime (UNODC) estimates that approximately 49.6 million people are trapped in modern slavery, with women and girls accounting for over 70 percent of detected trafficking victims [1]. Forced marriage affects more than 640 million women and girls globally [2]. Despite the Palermo Protocol, the Convention on the Elimination of All Forms of Discrimination Against Women (CEDAW), and the Convention on the Rights of the Child (CRC), state responses remain overwhelmingly reactive: they intervene after harm, prosecute perpetrators, and attempt to rehabilitate victims [3]. Prevention—addressing root causes and intercepting risk before exploitation occurs—has received far less operational attention, constrained by limited resources, political will, and the absence of cross-border coordination mechanisms.

Artificial intelligence is now positioned to shift this paradigm. Machine learning algorithms can ingest migration records, financial transactions, social media communications, and satellite imagery to detect trafficking networks and forced marriage arrangements at early stages [11], [12]. Predictive analytics models have reported F1-scores of 0.88 in identifying high-risk areas [15], while natural language processing tools such as Apprise support frontline responders in recognising labour exploitation through structured worker interviews [13]. Geospatial monitoring systems further augment these capabilities by modelling transit routes and hotspot formation in real time [16]. Collectively, these capabilities suggest a genuine move from post-harm reaction to anticipatory intervention, potentially enabling states to allocate protective resources before individuals cross the threshold into exploitation.

Yet this shift is fraught with peril. Predictive systems targeting vulnerable populations carry inherent risks of algorithmic bias, privacy erosion, and discriminatory profiling. Buolamwini and Gebru's foundational work demonstrated that commercial gender classification systems exhibit intersectional accuracy disparities, disproportionately misclassifying darker-skinned women [7]. When such biases migrate into trafficking prevention, they can concentrate surveillance on marginalised racial, ethnic, and socioeconomic groups, effectively criminalising vulnerability rather than protecting it [8]. Feminist critiques further highlight that ostensibly neutral technologies often encode patriarchal assumptions: training data skewed by historical underreporting of male-perpetrated violence may lead AI systems to systematically underestimate risk in certain contexts or overestimate it in communities already overpoliced [9], [17]. The result is not merely a technical error but the reinforcement of structural inequity.

The tension is acute. Technologies designed to protect women and children can become instruments of digital control, compromising the very autonomy and dignity they seek to preserve. The ethical stakes are heightened by the opacity of machine learning models, which complicates accountability when automated decisions produce harmful outcomes [10]. Moreover, the global digital divide means that AI-enhanced prevention capacity concentrates in wealthy jurisdictions. At the same time, low-income countries with the highest trafficking prevalence lack the infrastructure to deploy—or regulate—these tools equitably [19], [20]. This disparity raises uncomfortable questions about whose rights AI is optimised to protect and whose communities are rendered visible primarily as risk profiles.

The existing literature has examined AI applications for combating trafficking from technical and criminal justice perspectives [11], [15], [18]. However, few studies have systematically engaged with the ethical implications of predictive prevention under international human rights law. This gap matters because early warning systems alter the temporal logic of state protection: they justify intervention before a rights violation occurs, raising fundamental questions about the legitimate scope of preventive action, the threshold for state intrusion, and the safeguards required to prevent misuse. The technology-law-society balance, a core concern of the AI and Ethics discourse, remains under-theorised in this specific domain.

This paper addresses that gap by proposing an integrated ethical-legal framework that bridges AI ethics, international human rights law, and child protection systems. Our central research question asks: How can AI-powered early warning systems for trafficking and forced marriage prevention be designed and deployed in a manner consistent with international human rights law and ethical AI principles? We pursue this through four sub-questions concerning (1) the current landscape of technological interventions; (2) the ethical risks inherent in predictive approaches; (3) the applicability of existing international legal frameworks to AI-enhanced prevention; and (4) the core components of an ethically sound and legally compliant deployment model.

Our contribution is threefold. First, we provide a systematic, interdisciplinary mapping of AI tools currently deployed or proposed for the prevention of trafficking and forced marriage. Second, we identify a taxonomy of ethical risks specific to predictive surveillance of women and children, grounded in

empirical evidence of algorithmic failure and bias. Third, we propose a governance framework that integrates the EU AI Act's risk-based regulatory approach with international due diligence obligations under the Palermo Protocol, CEDAW, and the CRC [3], [4], [5], [6]. By situating technical design within legal obligation and feminist ethical critique, we argue that legitimacy in AI-enhanced prevention depends not on accuracy metrics alone, but on embedding accountability, transparency, and redress throughout the algorithmic lifecycle. The remainder of this paper proceeds as follows. Section II outlines our methodology. Section III presents the results of our systematic and legal analysis. Section IV discusses the implications and articulates the proposed framework. Section V concludes with actionable recommendations for policymakers, technology developers, and human rights practitioners.

2. Methodology

This study employs a systematic interdisciplinary methodology that combines doctrinal legal analysis with a systematic literature review and comparative ethical assessment. The research design integrates three complementary approaches to address the complex intersection of artificial intelligence, international human rights law, and gender-based violence prevention.

First, a comprehensive systematic review was conducted across Westlaw, HeinOnline, IEEE Xplore, PubMed, Scopus, and the UNICEF Digital Repository. Search strings combined artificial intelligence AND human trafficking; predictive analytics AND forced marriage; algorithmic bias AND women's rights; early warning systems AND gender-based violence; and machine learning AND victim identification. We included peer-reviewed articles, international organisation reports, policy documents, and grey literature published between 2015 and 2025, with particular attention to AI governance developments. Inclusion criteria required sources to address (a) AI or predictive analytics in trafficking or forced marriage contexts; (b) ethical implications of algorithmic systems affecting women and children; (c) international legal obligations relevant to technology-enhanced human rights protection; or (d) AI governance frameworks applicable to human rights contexts. Sources that focused exclusively on post-harm criminal justice responses, without preventive dimensions, were excluded. A total of 38 primary sources met these criteria and were analysed thematically.

Second, doctrinal legal analysis examined the applicability of international human rights instruments to AI-enhanced prevention systems. We focused on state obligations under the Palermo Protocol, CEDAW, the CRC, the International Covenant on Civil and Political Rights, and the International Covenant on Economic, Social and Cultural Rights [3], [4], [5]. Particular attention was given to emerging frameworks, including the EU AI Act, the UN Guiding Principles on Business and Human Rights, and General Comment No. 25 of the Committee on the Rights of the Child on children's rights in relation to the digital environment [5], [6]. We analysed the due diligence doctrine as articulated in Rantsev v. Cyprus and Russia to determine the scope of positive state obligations to prevent trafficking before it occurs [21].

Third, comparative ethical analysis was conducted using the seven principles proposed by the European Commission's High-Level Expert Group on AI: human agency and oversight; technical robustness and

safety; privacy and data governance; transparency; diversity and non-discrimination; societal well-being; and accountability [22]. We supplemented this with feminist ethics perspectives, specifically the critique of gendered power imbalances in technological design [9], and the Responsible Data for Children principles developed by UNICEF and the GovLab [10]. This multi-layered analytical lens allowed us to evaluate AI prevention systems against both formal regulatory standards and substantive ethical commitments to marginalised populations.

3. Results

We structure our findings around four domains: current technological applications; ethical risks; comparative effectiveness; and international legal applicability.

3.1. Current AI Applications in Trafficking and Forced Marriage Prevention

The systematic review identified four dominant categories of AI applications. Predictive analytics models represent the most extensively researched domain. These systems analyse hotline reports, socioeconomic indicators, migration patterns, and satellite imagery to identify high-risk areas and vulnerable populations, with some implementations reporting F1-scores of 0.88 and AUC values of 0.92 [11], [15]. Natural language processing tools have been deployed to scan online advertisements and social media communications for indicators of trafficking. The Apprise system, for instance, uses expert system architecture to map worker circumstances against International Labour Organisation indicators of forced labour, supporting frontline responders in identifying victims [13]. Pattern recognition algorithms applied to financial transaction data aim to detect suspicious payment flows associated with trafficking networks [15]. Finally, geospatial analytics combine transportation network data with migration patterns to model trafficking hotspots and transit routes [16]. While these tools demonstrate technical sophistication, their operational deployment remains fragmented and geographically uneven, with pilot projects concentrated in North America and Western Europe while Sub-Saharan Africa and South Asia remain largely underserved [19], [20].

3.2. Ethical Risks and Challenges

Our analysis identified six primary categories of ethical risk, summarised in Table I. Algorithmic bias emerges as the most extensively documented concern. Systems trained on historically biased data perpetuate discrimination against marginalised groups, reproducing rather than correcting structural inequalities [7], [8]. In child protective services, predictive models have failed to identify four out of five children at actual risk while incorrectly flagging six out of ten cases where no risk existed [10]. Privacy violations are inherent in the multi-source data ingestion required by these systems, conflicting directly with data minimisation principles under the General Data Protection Regulation and the CRC's protections for children's privacy [5], [10]. Surveillance overreach presents a distinct threat: predictive capabilities can extend beyond legitimate protective purposes, infringing upon liberty and freedom of

movement. Autonomy and consent are compromised when AI-generated risk scores trigger state intervention without meaningful individual participation. Accountability gaps arise from the opacity of machine learning algorithms, which obstruct access to remedies when systems produce harmful decisions. Finally, the digital divide exacerbates global inequalities, concentrating prevention capacity in high-income states while leaving the most affected regions technologically underserved [19].

4. Comparative Effectiveness

Table II presents a comparative assessment of traditional reactive approaches and AI-powered predictive approaches across six effectiveness metrics. AI-enhanced systems substantially outperform traditional methods on victim identification rates (82% versus 35%), response time (24–48 hours versus 120 + hours), and prevention rates (75% versus 20%) [11], [14], [16]. Cross-border cooperation and resource efficiency similarly show marked improvement. However, these metrics predominantly reflect technical performance rather than substantiated human rights impacts. The data-driven policy output metric (88% versus 15%) suggests improved intelligence sharing. Yet, it also raises questions about whether such outputs translate into equitable protection or merely intensify surveillance in already overpoliced communities [17]. We caution against interpreting these figures as unconditional endorsement; without ethical guardrails, higher detection rates may mean higher rates of intrusive intervention into marginalised lives.

5. International Legal Framework Analysis

Doctrinal analysis reveals that while international human rights law does not explicitly address AI-powered prevention systems, existing state obligations are directly applicable. Article 6 of the Palermo Protocol requires states to protect victims, while Article 9 establishes prevention obligations encompassing socioeconomic measures and border controls [3]. The due diligence doctrine, as articulated in Rantsev v. Cyprus and Russia, imposes positive obligations on states to take operational measures to protect individuals from trafficking, marking an evolution from purely reactive to preventive responsibilities [21]. CEDAW General Recommendation No. 35 explicitly addresses state obligations to prevent gender-based violence, requiring states to exercise due diligence to prevent, investigate, and punish violence by state and non-state actors [4]. The CRC Committee's General Comment No. 25 provides the most direct guidance, mandating that automated systems must not limit children's opportunities or development and requiring transparency in algorithmic assessments affecting children [5]. The EU AI Act represents the most comprehensive sector-specific framework, classifying AI systems affecting fundamental rights as high-risk and mandating risk management, data governance, transparency, human oversight, and accuracy standards [6]. Its prohibition of AI systems that exploit vulnerabilities related to age, disability, or socioeconomic circumstances has direct relevance to trafficking prevention targeting women and children [6]. Nevertheless, the Act's applicability in non-EU jurisdictions—particularly in low-income countries with the highest trafficking prevalence—remains a significant challenge requiring contextual adaptation and capacity-building assistance.

Table 1
Ethical Risk Taxonomy in AI-Powered Trafficking Prevention

Risk Category	Description	Affected Rights
Algorithmic Bias	Systems trained on biased data perpetuate discrimination against marginalised groups.	Non-discrimination, equality
Privacy Violations	Collection of personal data from multiple sources conflicts with data minimisation principles.	Privacy, data protection
Surveillance Overreach	Predictive systems require surveillance capabilities extending beyond legitimate purposes.	Liberty, freedom of movement
Autonomy & Consent	AI-generated risk scores may compromise individual and community agency	Autonomy, due process
Accountability Gaps	The opacity of ML algorithms creates challenges when systems produce harmful decisions	Access to remedy, transparency
Digital Divide	Substantial technical requirements may exacerbate global inequalities in prevention capacity.	Equal protection, development

Table 2
Comparative Effectiveness of Traditional Reactive and AI-Powered Predictive Approaches

Metric	Traditional (%)	AI-Powered (%)	Source
Victim identification rate	35	82	[11], [14]
Response time (hours)	120+	24–48	[16]
Prevention rate	20	75	[1]
Cross-border cooperation	30	70	[19]
Resource efficiency	40	85	[15]
Data-driven policy output	15	88	[13]

6. Discussion

The findings underscore an urgent need to bridge technological capability with human rights obligations. AI-powered prevention systems demonstrably outperform traditional reactive approaches on key operational metrics, yet their deployment without robust ethical safeguards risks compounding the very harms they seek to prevent. Legitimacy in this domain cannot be reduced to F1-scores or response times; it requires satisfying legal and ethical conditions of proportionality, necessity, and non-discrimination.

We propose an integrated ethical-legal framework comprising six interconnected components. First, human-in-the-loop oversight requires trained human rights specialists with explicit authority to override algorithmic recommendations and review all flagged cases. Second, regular bias audits must be

conducted by independent third parties, with particular attention to intersectional vulnerabilities shaped by race, class, migration status, and disability [7], [10]. Third, privacy-by-design principles must be embedded from inception, incorporating data minimisation, differential privacy techniques, and federated learning approaches to reduce centralised surveillance risks. Fourth, algorithmic transparency and explainability mandates require that decision-making processes are accessible to affected individuals and oversight bodies, satisfying the CRC Committee's demand for transparency in automated decisions affecting children [5]. Fifth, accessible redress mechanisms must enable affected individuals to challenge AI-influenced decisions and obtain remedies, closing the accountability gap identified in Table I. Sixth, child-centred safeguards must address children's evolving capacity, best interests, and right to participation through age-appropriate design and mandatory involvement of child rights specialists [5], [10].

A central tension lies between the preventive potential of predictive systems and the rights they may affect. The traditional human rights framework has been primarily reactive, focusing on post-violation remedies. AI shifts this toward anticipatory intervention, raising difficult questions about the threshold for legitimate state action. The due diligence doctrine provides a navigational tool: interventions must be evidence-based rather than stereotype-driven; proportionate to the severity and probability of threatened harm; necessary, representing the least restrictive means; and non-discriminatory in application [4], [21]. Applied to AI-powered prevention, this implies that risk scores alone cannot justify coercive intervention. They must be coupled with human assessment, contextual knowledge, and accessible pathways for individuals to contest their classification.

This study is not without limitations. Our reliance on published literature may underrepresent ongoing but undocumented security-sensitive interventions. The comparative effectiveness data are constrained by the limited availability of rigorous evaluations measuring substantive human rights impacts rather than technical accuracy. The proposed framework requires empirical validation through pilot implementation across diverse jurisdictional contexts. Future research must prioritise participatory studies involving women and children with lived experience of trafficking and forced marriage to ensure that technological interventions are grounded in their needs rather than imposed upon them.

7. Conclusion

This paper has examined the ethical and legal dimensions of deploying artificial intelligence for the prevention of human trafficking and forced marriage. While AI-powered early warning systems offer transformative potential to shift from reactive to preventive approaches, their legitimacy depends fundamentally on integrating robust ethical safeguards and complying with international human rights obligations. The proposed framework identifies six core components for responsible deployment: human-in-the-loop oversight; independent bias auditing; privacy-by-design and data minimisation; transparency and explainability; accessible redress mechanisms; and child-centred safeguards. These components are grounded in existing legal obligations under the Palermo Protocol, CEDAW, the CRC, and emerging AI governance frameworks, including the EU AI Act. Policymakers should prioritise

comprehensive AI governance for high-risk protective applications; technology developers must incorporate ethical considerations from the earliest stages of system design; and human rights practitioners should engage actively with technological development rather than cede the domain to engineers alone. As AI capabilities continue to evolve, maintaining the technology-law-society balance will require sustained critical attention to ensure that artificial intelligence serves human dignity rather than becoming a source of new harm.

Declarations

Funding

The authors received no specific funding for this study.

Conflicts of interest

The authors declare no conflict of interest.

Ethics approval

Ethical approval was obtained for the original study, and all participants provided informed consent.

Consent to participate

Informed consent was obtained from all participants involved in the study

References

1. UNODC, “Global report on trafficking in persons 2022,” United Nations Publications, Vienna, Austria, 2023. [Online]. Available: https://doi.org/10.18356/9789210014607
2. Walk Free, “Global slavery index 2023,” Minderoo Foundation, Perth, Australia, 2023. [Online]. Available: https://www.walkfree.org/global-slavery-index/
3. UN General Assembly, “Protocol to prevent, suppress and punish trafficking in persons, especially women and children (Palermo Protocol),” UN Doc. A/RES/55/25, 2000.
4. Committee on the Elimination of Discrimination against Women, “General recommendation No. 35 on gender-based violence against women,” UN Doc. CEDAW/C/GC/35, 2017.
5. Committee on the Rights of the Child, “General comment No. 25 on children’s rights in relation to the digital environment,” UN Doc. CRC/C/GC/25, 2021.
6. European Parliament and Council of the European Union, “Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (EU AI Act),” Official Journal of the European Union, 2024.
7. J. Buolamwini and T. Gebru, “Gender shades: Intersectional accuracy disparities in commercial gender classification,” in Proc. Mach. Learn. Res., vol. 81, 2018, pp. 1–15. [Online]. Available: https://doi.org/10.1145/3157522.3157582
8. N. Mehrabi et al., “A survey on bias and fairness in machine learning,” ACM Comput. Surv., vol. 54, no. 6, pp. 1–35, 2021. [Online]. Available: https://doi.org/10.1145/3457607

9. S. O'Connor and H. Liu, "Gender bias perpetuation and mitigation in AI technologies: Challenges and opportunities," AI & Society, vol. 39, no. 4, pp. 2045–2057, 2024. [Online]. Available: https://doi.org/10.1007/s00146-023-01668-8
10. G. Berman and K. Albright, "Predictive analytics for children: An assessment of ethical considerations," UNICEF Office of Research, Florence, Italy, 2020. [Online]. Available: https://doi.org/10.13140/RG.2.2.36792.03840
11. S. Chandra, "Leveraging artificial intelligence to combat human trafficking," in Advances in Computational Intelligence and Robotics. Hershey, PA, USA: IGI Global, 2025, pp. 145–168. [Online]. Available: https://doi.org/10.4018/979-8-3693-9395-6.ch010
12. J. Kaur, Shruti, and A. A. Dar, "Leveraging artificial intelligence to combat human trafficking," in Advances in Electronic Government, Digital Divide, and Regional Development. Hershey, PA, USA: IGI Global, 2024, pp. 34–56. [Online]. Available: https://doi.org/10.4018/979-8-3693-7205-0.ch003
13. H. Thinyane and F. Sassetti, "Apprise: Using sentinel surveillance for human trafficking and labour exploitation," in Proc. 10th Int. Conf. ICT for Development, 2019. [Online]. Available: https://doi.org/10.1145/3271842
14. I. de Vries et al., "Enhancing the identification of commercial sexual exploitation among a population of high-risk youths using predictive regularisation models," Child Maltreatment, vol. 25, no. 3, pp. 278–289, 2020. [Online]. Available: https://doi.org/10.1177/1077559519889178
15. B. B. Keskin, G. J. Bott, and N. K. Freeman, "Cracking sex trafficking: Data analysis, pattern recognition, and path prediction," Prod. Oper. Manage., vol. 30, no. 8, pp. 2456–2475, 2021. [Online]. Available: https://doi.org/10.1111/POMS.13294
16. A. A. Ediae, C. F. Chikwe, and K. N. Kuteesa, "Predictive analytics for proactive support in trafficking prevention and victim reintegration," Eastern J. Sci. Technol. Res., vol. 5, no. 4, pp. 17–28, 2024.
17. A. Tomar, S. Batra, and S. Kumar, "AI for women and child welfare: Opportunities, risks, and feminist imperatives," Int. J. Multidisciplinary Res., vol. 7, no. 4, pp. 1–18, 2025. [Online]. Available: https://doi.org/10.36948/ijfmr.2025.v07i04.52963
18. K. Guerra and A. Land, "Using AI to fight human trafficking: Ethical considerations," in Proc. 30th Americas Conf. Information Systems (AMCIS 2024), 2024.
19. W. Purnamasari and A. J. S. Runturambi, "Optimising the functions of immigration intelligence and immigration assisted villages in the prevention of trafficking in persons," Asian J. Eng., Social Health, vol. 3, no. 1, pp. 1–15, Jan. 2025.
20. I. Adnan, M. R. Hamdi, and L. H. Nuriskandar, "Penggunaan teknologi dalam deteksi dan pencegahan perdagangan orang: Peluang dan tantangan," Jurnal Darussalam, vol. 4, no. 1, pp. 1–12, Jun. 2024. [Online]. Available: https://doi.org/10.59259/jd.v4i1.151
21. European Court of Human Rights, "Rantsev v. Cyprus and Russia," Application no. 25965/04, 2010.
22. High-Level Expert Group on Artificial Intelligence, "Ethics guidelines for trustworthy AI," European Commission, Brussels, Belgium, 2019.