Ask before you Build: Rethinking AI-for-Good in Human Trafficking Interventions

Pratheeksha Nair
pratheeksha.nair@mail.mcgill.ca
McGill University, Mila-Quebec AI
Institute
Montreal, Canada

Gabriel Lefebvre
gabriel.lefebvre@mail.mcgill.ca
McGill University
Montreal, Canada

Sophia Garrel
sophiagarrel@gmail.com
Universite de Montreal, Mila-Quebec
AI Institute
Montreal, Canada

Maryam Molamohammadi
Mila-Quebec AI Institute
Montreal, Canada

Reihaneh Rabbany
rrabba@cs.mcgill.ca
McGill University, Mila-Quebec AI
Institute
Montreal, Canada

Abstract

AI-for-good initiatives often rely on the assumption that technical interventions can resolve complex social problems. In the context of human trafficking (HT), such techno-solutionism risks oversimplifying exploitation, reinforcing power imbalances, and causing harm to the very communities AI claims to support. In this paper, we introduce the Radical Questioning (RQ) framework as a five-step, pre-project ethical assessment tool to critically evaluate whether AI should be built at all—especially in domains involving marginalized populations and entrenched systemic injustice. RQ does not replace principles-based ethics but precedes it, offering an upstream, deliberative space to confront assumptions, map power, and consider harms before design. Using a case study in AI for HT, we demonstrate how RQ reveals overlooked socio-cultural complexities and guides us away from surveillance-based interventions toward survivor-empowerment tools. While developed in the context of HT, RQ's five-step structure can generalize to other domains—though the specific questions must be contextual. This paper situates RQ within a broader AI ethics philosophy that challenges instrumentalist norms and centers relational, reflexive responsibility.

CCS Concepts

• **Social and professional topics** → *Computer crime*; • **Computing methodologies** → **Philosophical/theoretical foundations**

Authors' Contact Information: Pratheeksha Nair, pratheeksha.nair@mail.mcgill.ca, McGill University, Mila-Quebec AI Institute, Montreal, Canada; Gabriel Lefebvre, gabriel.lefebvre@mail.mcgill.ca, McGill University, Montreal, Canada; Sophia Garrel, sophiagarrel@gmail.com, Universite de Montreal, Mila-Quebec AI Institute, Montreal, Canada; Maryam Molamohammadi, Mila-Quebec AI Institute, Montreal, Canada; Reihaneh Rabbany, rrabba@cs.mcgill.ca, McGill University, Mila-Quebec AI Institute, Montreal, Canada.

of artificial intelligence; • **General and reference** → *Computing standards, RFCs and guidelines.*

Keywords

AI Ethics, Responsible AI development, Radical Questioning

1 Introduction

In recent years, the deployment of artificial intelligence (AI) in social good contexts has surged, driven by the belief that technology can offer effective solutions to complex societal challenges [22]. Most AI-for-good projects inherit a techno-solutionist mindset [23, 32, 41, 44]—believing that complex social problems are reducible to datasets and solvable via automation. In domains like human trafficking (HT), this framing is not only inadequate but dangerous: it reifies harmful narratives, empowers surveillance actors, and silences those directly affected.

Recent AI interventions targeting HT have largely focused on crime-sleuthing applications, including machine learning classifiers for escort ads [1, 15, 29, 31, 37, 48], computer vision systems for identifying trafficking indicators in imagery [47] and geotags [3], and network-mapping tools used by law enforcement and NGOs to detect trafficking "hotspots" [2]. These tools typically frame trafficking as a data problem—one to be solved through pattern recognition and predictive modeling. However, critics note that such approaches rest on narrow assumptions, prioritize enforcement over survivor well-being, and frequently ignore systemic root causes [13, 35, 36]. Studies have shown that these models can perpetuate racialized and gendered biases, misclassify consensual sex work as trafficking, and ultimately reinforce surveillance and carceral systems [10]. We argue that these shortcomings calls for upstream, human-centered frameworks that interrogate the ethics of intervention before technical development begins.

Situating this upstream intervention within the existing ethical frameworks [19, 28, 30] — these center on principles like transparency and fairness — are reactive and compliance-oriented, and

often implemented only after technical design has begun. These seldom ask a more fundamental question: Should we be building this at all?

In this paper, we introduce Radical Questioning (RQ) as an upstream, pre-design ethics framework that foregrounds this very question. Developed through interdisciplinary collaboration in the HT context, RQ guides AI developers through critical reflection about the motivations, implications, and legitimacy of their proposed intervention. Unlike participatory design [8, 13, 23] or risk assessments [38], RQ is agnostic to outcomes and rooted in philosophical deliberation [7, 17, 20] — it does not prescribe what should be done, but how to interrogate why we believe something ought to be done in the first place.

RQ's novelty lies in both its structural role—intervening before AI design—and its application to the hyper-contested, ethically fraught domain of HT, which provides a particularly acute example of how AI can exacerbate harm under the guise of social good. While developed in the HT context, RQ's structure is generalizable to other settings where ethical complexity precedes technical feasibility. Moreover, we call for a paradigm shift in AI ethics: to normalize pre-project ethical assessment as a core component of responsible AI practice, particularly in domains marked by contested definitions, systemic harms, and stakeholder asymmetries.

2 Related Works

Our framework builds on a growing body of research that critiques techno-solutionism and advocates for context-sensitive, justice-oriented approaches to AI design. Refusal-based frameworks such as Studying Up [5] and the broader theory of data refusal [4] challenge dominant problem framings by interrogating harmful assumptions and the positionality of power. Similarly, Design Justice [11] and the heterogeneity framework [34] emphasize centering marginalized voices and accounting for intersections of culture, law, and infrastructure. While RQ shares these commitments to power analysis and critical reflection, it extends them into a structured, iterative, and pre-design methodology that explicitly asks whether an AI system should be built at all.

Other scholars call for re-centering AI ethics on foundational moral questions [22, 23], or critique how AI amplifies state surveillance and racialized carceral logics in the context of trafficking [33, 36]. Complementary tools like RED (Rapid Ethical Deliberation) [45] provide procedural or risk-centered guidance [40] for high-stakes domains but typically assume a system is already in development.

Deliberation-focused approaches like the Situate AI Guidebook [24] and participatory frameworks [14] support structured stakeholder engagement, while HCI contributions—such as trauma-informed computing [25], survivor-centered justice [39], and digital-safety research protocols [6]—offer critical guidance for working with vulnerable populations. However, many of these operate downstream of the initial decision to build and rarely offer mechanisms to pause or reframe interventions entirely. RQ complements and deepens these efforts by introducing open-ended, upstream deliberation that challenges the necessity and legitimacy of technical intervention at its root.

In the HT domain, prior work has identified risks such as dataset bias, privacy violations, and the harms of over-policing, especially when AI tools are deployed without attention to socio-cultural complexity [8, 13]. The RQ framework addresses this gap by offering a pre-design process that surfaces these concerns early and critically, enabling more ethically sound decisions before systems are built. Rather than relying on fixed stakeholder roles or prescriptive checklists, RQ creates reflective space for foundational questioning—making it compatible with but not dependent on institutional or participatory structures. Within HT, it uniquely reshapes design trajectories by grounding intervention in relational ethics and moral deliberation.

3 Radical Question Framework

We introduce Radical Questioning (RQ) as a pre-design ethics framework developed through its application in the human trafficking (HT) domain. RQ refers to the practice of interrogating foundational assumptions that shape how problems are defined and solutions are pursued. Rather than asking how to build better AI tools for combating HT, RQ begins by asking why AI is the appropriate response at all—and who benefits or is harmed by its deployment. Questions such as "What does justice mean in this context?" and "Whose definitions are being used, and why?" invite critical reflection on the normative, political, and institutional stakes of AI-for-good projects.

The Radical Questioning (RQ) framework emerged through sustained engagement with survivor-led organizations and interdisciplinary collaborators, grounded in the legal and social complexities of the human trafficking (HT) domain. While context-specific in origin, RQ's five-step process is broadly applicable to other ethically complex domains—so long as its questions are rooted in the specific realities of those fields. Designed for AI developers, researchers, policymakers, and interdisciplinary teams, RQ is especially valuable in high-stakes contexts where definitions are contested and harm is unevenly distributed. It guides not implementation, but reflection on whether an AI intervention should proceed at all.

Although not a participatory design method, RQ is grounded in relational accountability: its questions were co-shaped by individuals differently situated in relation to harm, power, and intervention. Rather than a checklist, RQ is a deliberative practice that invites open-ended inquiry, iterative reflection, and deep stakeholder engagement to resist premature ethical closure.

Below, we present RQ's five steps, accompanied by example questions developed and refined through our HT case study. Gray boxes in the following section highlight actual questions raised during the design process to illustrate how RQ unfolds in practice. We also present RQ diagrammatically in Figure 1.

3.1 Step 1: Define the Scope of the Problem

This step asks not what the technical problem is, but what the social issue being addressed actually means—and who gets to define it. In our HT case, the initial framing was: "How can we detect trafficking online using escort advertisements?" Yet this presumes that trafficking is legible to machine learning, that online ads are reliable indicators, and that detection is the right goal. By asking: Why are we framing HT in this way? Who benefits from this definition? we learned that much of the literature and tooling assumes a fixed, binary notion of exploitation.
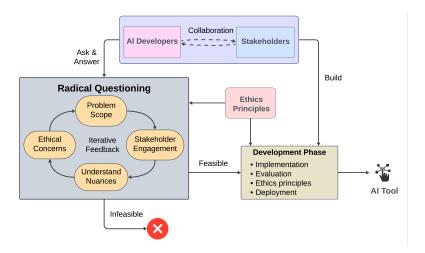
Figure 1: The proposed RQ framework is based on asking and answering radical questions through deliberative communication and collaboration between AI developers and stakeholders before development. If deemed ethically feasible, the tool is developed and if not, it is terminated.

HT laws in Canada, for example, are often invoked in ways that conflate sex work with exploitation, and are disproportionately used against migrant sex workers [10, 26]. Moreover, assumptions such as "pimps write the ads" [9, 12] ignore survivor agency and may lead to harmful interventions. Such assumptions erase the agency of sex workers and legitimizes surveillance-based interventions that can put marginalized communities at further risk. This shifted our framing from "detection of trafficking" to "support for survivors."

- What is human trafficking and why is it problematic?
- Why are we defining the problem in this particular way, and who benefits from this framing?
- Is this a problem that can be alleviated using AI? Is there a demand for solving this problem and who has raised the demand?
- Are there particular problems within the HT domain that we can focus on? What sort of resources are required for solving these problems?

3.2 Step 2: Identify Stakeholders

AI development often privileges institutional stakeholders—such as law enforcement or funders—while marginalizing the perspectives of those most affected. We asked: Who will be impacted by this tool? Who owns it? Are our designs centered around who holds power or the proximity to potential harm?

In the HT ecosystem, stakeholder perspectives diverge drastically. Law enforcement views tech as a tool for surveillance and arrest [16, 48]; NGOs often focus on victim identification [27, 42]; sex worker organizations advocate for harm reduction and autonomy [10]. Survivors themselves express complex, sometimes ambivalent relationships to justice and intervention [18, 46]. We found that involving survivors directly—not just via proxies—radically

altered the design direction. This engagement also revealed logistical challenges: many survivor groups had justifiable distrust in institutional actors, and sustained trust-building was necessary.

- Who will be the end user of our tool? Who will own the tool and ensure its proper use?
- How do the different stakeholders implicated in the project understand the function and limits of the criminal law/and technology?
- Are we privileging certain stakeholders (e.g., law enforcement) over others (e.g., survivors), and what are the implications?
- Have we involved those directly affected by the tool, including marginalized voices, in meaningful ways?
- Do we have sufficient resources and expertise to engage meaningfully with multiple stakeholders?

3.3 Step 3: Understand Contextual Nuance

Even when stakeholders are engaged, ethical blind spots remain if practitioners don't grapple with domain complexity. In HT, notions of justice, consent, and exploitation are far from universal. Some survivors do not want to press charges or be rescued; others fear the criminal justice system more than their traffickers. Even identifying "risk factors" (e.g., ethnic markers such as "black", "asian", extreme services like "BDSM", "sodomy"', sexual orientation indicators like "queer" or "trans", or language cues like poor English indicating foreigners) [21] risks reinforcing racial and gendered surveillance.

A particularly salient example is the "chilling effect" [43, 50] on sex workers caused by AI-based ad monitoring: fear of criminalization leads to self-censorship and economic insecurity. RQ helped us avoid replicating such harm by reframing success not as catching traffickers, but as supporting survivors' pursuit of justice on their terms.

- Have we uncovered the underlying complexities of the problem? What do these complexities mean for each stakeholder?
- How do different stakeholders perceive the problem we aim to address and do they agree with our approach of addressing it?
- How do we measure the success of the tool, and who decides what success looks like?
- Is our metric of success causing unintended harms to any stakeholders?

- Are there systems in place to obtain continuous feedback from stakeholders/survivors? How to be mindful of the risks of retraumatization?
- Are we actively considering critiques that contradict our initial goals, or are we selectively responding to feedback that supports our pre-existing views?
- Are we truly acting on the feedback, or simply acknowledging it to maintain the appearance of responsiveness?
- Are there stakeholders whose feedback we are dismissing because it challenges the feasibility or goals of the project?

3.4 Step 4: Map Ethical Concerns

This step surfaces questions around accountability, privacy, fairness, legitimacy, and incentives. We asked: What does fairness mean in this domain? Who defines accountability? We found that ground truths for "exploitation" are based on legal and social interpretations that evolve—and that AI tools trained on them can reinforce biased narratives [21, 36]. Equally, developers often escape accountability, while the tool becomes a mechanism for state power. RQ helped surface the disconnect between legal structures and social legitimacy. Additionally, in HT, legal compliance alone is insufficient for legitimacy [26]. For instance, privacy laws may permit certain uses of "public" data, but affected communities may still experience real harm. For example, an AI tool that flags "suspicious" ads can generate false positives, prompting police surveillance of consensual sex workers and retraumatizing victims.

- What measures are in place to assess the tool's fairness, accuracy, and social impact? Who decides what is fair and accurate? Which stakeholders were involved in discussing these metrics?
- How to choose who needs to take accountability for the tool? What does accountability mean to those involved in the domain?
- Is simple conformity to law and constitutional requirements a sufficient bases for legitimacy of the initiative?
- What does privacy mean for those affected by the problem and the tool?
- What is the composition of the team behind the tool and what are their motives and incentives?

3.5 Step 5: Iterate with Feedback

Finally, RQ requires continuous stakeholder input, especially from vulnerable groups. But feedback is not merely about checking boxes—it must be deliberative, responsive, and open to halting a project entirely. We asked: Are we truly integrating critique? Are we willing to stop if harm outweighs benefit?

Our team participated in trauma-informed workshops with engaged with survivor-led organizations, while remaining attentive to the risks of retraumatization [49]. We built an advisory board including survivors to oversee the tool's development and function. Their feedback prompted design pivots—including reducing automation, foregrounding consent, and focusing on evidence preservation (not detection). Without this feedback, the project might have reinforced the very systems it sought to challenge.

4 Challenges and Limitations

While RQ proved valuable in this project, we also encountered several note-worthy challenges. 1) RQ is not prescriptive. Its strength lies in surfacing ethical dilemmas, but it does not dictate clear answers or technical implementation paths. This can be challenging for practitioners seeking actionable outcomes in time-constrained or product-driven environments. 2) The effectiveness of RQ hinges on genuine stakeholder engagement. In practice, building trust with affected communities—especially those historically marginalized or criminalized—requires time, resources, and institutional support. In our case, it took sustained outreach to survivor-led organizations before meaningful consultation could begin. In other contexts, such access may be limited or mediated through gatekeepers. 3) The outcomes of RQ are contingent on the positionality and openness of the development team. If teams are unwilling to alter project goals or confront uncomfortable truths, RQ risks becoming performative. Moreover, RQ requires developers to take on ethical responsibility themselves rather than outsourcing it to compliance checklists or institutional norms. 4) While the RQ process is transferable, the specific questions and concerns are not. Applying RQ to new domains will require re-grounding in context-specific histories, power dynamics, and social imaginaries. The framework is only as effective as the depth and sincerity of the questions asked.

5 Conclusions and Takeaways

Applying the Radical Questioning (RQ) framework to human trafficking (HT) fundamentally reshaped our project. What began as an AI-for-good intervention—automated detection of trafficking in online ads—evolved into a survivor-centered evidence management tool. This shift was not due to technical barriers, but ethical insight gained through interdisciplinary reflection and survivor engagement.

RQ prompted us to question core assumptions: Why prioritize detection? Who is helped—or harmed—by such tools? These inquiries revealed the risks of over-surveillance, false positives, and retraumatization. Engaging with survivors reoriented our goals from system enforcement to individual autonomy. We moved from detection to documentation; from surveillance to support. This pivot demanded humility and accountability. We established a survivor-led advisory board and adopted trauma-informed engagement practices. RQ revealed that building the "right" AI tool sometimes means not building the original one at all.

Though born from HT, RQ is domain-agnostic. Its five steps can guide AI projects in other high-stakes areas like child welfare or

predictive policing, where social complexity and contested ethics are common. However, its value lies not in universal questions but in enabling domain-specific deliberation. RQ is not a checklist—it's a scaffold for relational ethics, adaptable to differing legal, cultural, and moral contexts.

RQ repositions ethics as foundational—not auxiliary—to design. It challenges the dominant logic of "build fast, comply later," advocating instead for a stance of relational responsibility. In domains where harm is obscured by urgency and moral righteousness, such reflection is not optional—it's essential.

From our experience, we offer the following recommendations for researchers and practitioners considering the RQ approach: 1) Reframe ethical inquiry as an entry point, not an afterthought. Ethical reflection should begin before any AI design work—not after feasibility is established. 2) Institutionalize the option of not building. Ethical frameworks should allow space to walk away from a project if harms outweigh benefits. 3) Replace fixed checklists with iterative dialogue. Tools like RQ must remain flexible and open-ended, prioritizing critical reflection over formal compliance. 4) Prioritize relational accountability. Establish advisory structures or partnerships with affected communities that extend throughout the project lifecycle and not just during initial consultation. 5) Rethink success metrics. In sensitive domains, success may mean empowerment, harm reduction, or withdrawal—not optimization or scale. 6) Ultimately, RQ fosters a philosophy of deliberate ethics: one in which developers embrace uncertainty, remain open to critique, and recognize that responsible AI is not merely a matter of doing things right—but asking if they should be done at all.

References

[1] Hamidreza Alvari, Paulo Shakarian, and JE Kelly Snyder. 2017. Semi-supervised learning for detecting human trafficking. *Security Informatics* 6, 1 (2017), 1.

[2] Brittany Anthony and Contributing Survivor Authors. 2018. *On-Ramps, Intersections, and Exit Routes: A Roadmap for Systems and Industries to Prevent and Disrupt Human Trafficking – Social Media.* Technical Report. Polaris Project, Washington, DC. https://polarisproject.org/wp-content/uploads/2018/08/A-Roadmap-for-Systems-and-Industries-to-Prevent-and-Disrupt-Human-Trafficking-Social-Media.pdf

[3] Opeyemi Bamigbade, John Sheppard, and Mark Scanlon. 2024. Computer vision for multimedia geolocation in human trafficking investigation: A systematic literature review. *arXiv preprint arXiv:2402.15448* (2024).

[4] Chelsea Barabas. 2022. Refusal in data ethics: Re-imagining the code beneath the code of computation in the carceral state. *Engaging Science, Technology, and Society* 8, 2 (2022), 35–57.

[5] Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. 2020. Studying up: reorienting the study of algorithmic fairness around issues of power. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* 167–176.

[6] Rosanna Bellini, Emily Tseng, Noel Warford, Alaa Daffalla, Tara Matthews, Sunny Consolvo, Jill Palzkill Woelfer, Patrick Gage Kelley, Michelle L Mazurek, Dana Cuomo, et al. 2024. Sok: Safer digital-safety research involving at-risk users. In *2024 IEEE Symposium on Security and Privacy (SP).* IEEE, 635–654.

[7] R Benjamin. 2019. *Race after Technology: Abolitionist Tools for the New Jim Code.* Cambridge and Medford: Polity Press.

[8] Rasika Bhalerao. 2022. *Analyzing Harms of Online Platform and Policy Design.* New York University Tandon School of Engineering.

[9] Vanessa Bouché and Dana E Wittmer. 2015. Gendered diffusion on gendered issues: the case of human trafficking. *Journal of Public Policy* 35, 1 (2015), 1–33.

[10] Sydney Brown. 2024. Policing sex work online: sex workers' views on the risks and benefits of using AI to police online ads for sexual services. (2024).

[11] Sasha Costanza-Chock. 2020. *Design justice: Community-led practices to build the worlds we need.* The MIT Press.

[12] Sean M Crotty and Vanessa Bouché. 2018. The red-light network: Exploring the locational strategies of illicit massage businesses in Houston, Texas. *Papers in Applied Geography* 4, 2 (2018), 205–227.

[13] Julia Deeb-Swihart, Alex Endert, and Amy Bruckman. 2022. Ethical tensions in applications of ai for addressing human trafficking: A human rights perspective. *Proceedings of the ACM on human-computer interaction* 6, CSCW2 (2022), 1–29.

[14] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization.* 1–23.

[15] Artur Dubrawski, Kyle Miller, Matthew Barnes, Benedikt Boecking, and Emily Kennedy. 2015. Leveraging publicly available data to discern patterns of human-trafficking activity. *Journal of Human Trafficking* 1, 1 (2015), 65–85.

[16] Annalisa Enrile and Gabrielle Aquino-Adriatico. 2024. *Technology Innovations in Fighting Slavery and Human Trafficking.* Springer Nature Switzerland, Cham, 179–203. doi:10.1007/978-3-031-58614-9_10

[17] Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press.

[18] Amy Farrell, Meredith Dank, Ieke de Vries, Matthew Kafafian, Andrea Hughes, and Sarah Lockwood. 2019. Failing victims? Challenges of the police response to human trafficking. *Criminology & Public Policy* 18, 3 (2019), 649–673.

[19] Lajla Fetic, Torsten Fleischer, Paul Grünke, Thilo Hagendorf, Sebastian Hallensleben, Marc Hauer, Michael Herrmann, Rafaela Hillerbrand, Carla Hustedt, Christoph Hubig, et al. 2020. From Principles to Practice. An interdisciplinary framework to operationalise AI ethics. (2020).

[20] Nancy Fraser. 2009. Scales of Justice: Reimagining Political Space in a Globalizing World. *Polity* (2009).

[21] Luca Giommoni and Ruth Ikwu. 2021. Identifying human trafficking indicators in the UK online sex market. *Trends in Organized Crime* (2021), 1–24.

[22] Ben Green. 2019. Good" isn't good enough. In *Proceedings of the AI for Social Good workshop at NeurIPS*, Vol. 17.

[23] Jan-Christoph Heilinger. 2022. The ethics of AI ethics. A constructive critique. *Philosophy & Technology* 35, 3 (2022), 61.

[24] Anna Kawakami, Amanda Coston, Haiyi Zhu, Hoda Heidari, and Kenneth Holstein. 2024. The Situate AI Guidebook: Co-Designing a Toolkit to Support Multi-Stakeholder, Early-stage Deliberations Around Public Sector AI Proposals *(CHI '24).* Association for Computing Machinery, Article 749, 22 pages.

[25] Shannon Kelly, Eric Rodriguez, Stuart Blythe, and Ben Lauren. 2022. Trauma-Informed Scholarship in Digital Research and Design. *Methods and Methodologies for Research in Digital Writing and Rhetoric: Centering Positionality in Computers and Writing Scholarship* 2 (2022), 81–103.

[26] Renata A. Konrad, Kayse Lee Maass, Geri L. Dimas, and Andrew C. Trapp. 2023. Perspectives on how to conduct responsible anti-human trafficking research in operations and analytics. *European Journal of Operational Research* 309, 1 (2023), 319–329. doi:10.1016/j.ejor.2022.12.02

[27] Gabriel Lefebvre and Karim Benyekhlef. 2023. "Predictive policing in Canada" in "Artificial Intelligence and Administration of Criminal Justice". *International Review of Penal Law* 94 (2023).

[28] David Leslie. 2019. Understanding artificial intelligence ethics and safety. *arXiv preprint arXiv:1906.05684* (2019).

[29] Ruoting Li, Margaret Tobey, Maria E Mayorga, Sherrie Caltagirone, and Osman Y Özaltın. 2023. Detecting human trafficking: Automated classification of online customer reviews of massage businesses. *Manufacturing & Service Operations Management* 25, 3 (2023), 1051–1065.

[30] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI conference on human factors in computing systems.* 1–14.

[31] Anastasija Mensikova and Chris A Mattmann. 2018. Ensemble sentiment analysis to identify human trafficking in web data. In *Workshop on Graph Techniques for Adversarial Activity Analytics (GTA 2018), Marina Del Rey, CA, USA.* 5–9.

[32] Jacob Metcalf, Emanuel Moss, et al. 2019. Owning ethics: Corporate logics, silicon valley, and the institutionalization of ethics. *Social Research: An International Quarterly* 86, 2 (2019), 449–476.

[33] Sanja Milivojevic, Heather Moore, and Marie Segrave. 2020. Freeing the Modern Slaves, One Click at a Time: Theorising human trafficking, modern slavery, and technology. *Anti-trafficking review* 14 (2020), 16–32.

[34] Maryam Molamohammadi, Afaf Taïk, Nicolas Le Roux, and Golnoosh Farnadi. 2023. Unraveling the Interconnected Axes of Heterogeneity in Machine Learning for Democratic and Inclusive Advancements. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization.* 1–12.

[35] Jennifer Musto, Mitali Thakor, and Borislav Gerasimov. 2020. Between Hope and Hype: Critical evaluations of technology's role in anti-trafficking. *Anti-trafficking review* 14 (2020), 1–14.

[36] Jennifer Lynne Musto and Danah Boyd. 2014. The trafficking-technology nexus. *Social Politics* 21, 3 (2014), 461–483.

[37] Pratheeksha Nair, Javin Liu, Catalina Vajiac, Andreas Olligschlaeger, Duen Horng Chau, Mirela Cazzolato, Cara Jones, Christos Faloutsos, and Reihaneh Rabbany. 2024. T-NET: Weakly Supervised Graph Learning for Combatting Human Trafficking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38.

22276–22284.

[38] Ontario Human Rights Commission. n.d.. Human Rights and AI: Impact Assessment. https://www3.ohrc.on.ca/en/human-rights-ai-impact-assessment Accessed: 2025-01-21.

[39] Hawra Rabaan and Lynn Dombrowski. 2023. Survivor-Centered Transformative Justice: An Approach to Designing Alongside Domestic Violence Stakeholders in US Muslim Communities. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[40] Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Gianluca Stringhini, Thamar Solorio, Munmun De Choudhury, and Pamela J Wisniewski. 2021. A human-centered systematic literature review of the computational approaches for online sexual risk detection. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–38.

[41] Donna Riley. 2008. Engineering and social justice. In *Engineering and Social Justice*. Springer, 47–106.

[42] Rowena Rodrigues. 2020. Legal and human rights issues of AI: Gaps, challenges and vulnerabilities. *Journal of Responsible Technology* 4 (2020), 100005.

[43] Frederick Schauer. 1978. Fear, risk and the first amendment: Unraveling the chilling effect. *BUL rev.* 58 (1978), 685.

[44] Natasha Dow Schüll. 2013. The folly of technological solutionism: An interview with evgeny morozov.

[45] Marc Steen, Martijn Neef, and Tamar Schaap. 2021. A method for rapid ethical deliberation in research and innovation projects. *International Journal of Technoethics (IJT)* 12, 2 (2021), 72–85.

[46] Andrea Sterling and Emily van der Meulen. 2018. "We are not criminals": Sex work clients in Canada and the constitution of risk knowledge. *Canadian Journal of Law and Society/La Revue Canadienne Droit et Société* 33, 3 (2018), 291–308.

[47] Abby Stylianou, Richard Souvenir, and Robert Pless. 2019. TraffickCam: Explainable Image Matching For Sex Trafficking Investigations. *arXiv preprint arXiv:1910.03455* (2019).

[48] Edmund Tong, Amir Zadeh, Cara Jones, and Louis-Philippe Morency. 2017. Combating human trafficking with deep multimodal models. *arXiv preprint arXiv:1705.02735* (2017).

[49] Rachel Witkin and Katy Robjant. 2018. The Trauma-Informed Code of Conduct. *London: Helen Bamber Foundation* (2018).

[50] Hilary Young. 2023. Hansman v Neufeld: The Supreme Court of Canada protects counterspeech under anti-SLAPP law, but is it even defamatory? *Journal of Media Law* 15, 2 (2023), 125–139.