Reframing AI-for-Good: Radical Questioning in AI for Human Trafficking Interventions

Pratheeksha Nair[12], **Gabriel Lefebvre**[1],
Maryam Molamohammadi[2], **Sophia Garrel**[2], **Reihaneh Rabbany**[12]
[1]McGill University
[2]Mila-Quebec AI Institute
845 Sherbrooke St W, Montreal, Quebec
pratheeksha.nair@mail.mcgill.ca

Abstract

This paper introduces Radical Questioning (RQ), a structured, pre-design ethics framework developed to assess **whether** artificial intelligence (AI) should be applied to complex social problems rather than merely how. While much of responsible AI development focuses on aligning systems with principles such as fairness, transparency, and accountability, it often begins after the decision to build has already been made, implicitly treating the deployment of AI as a given rather than a question in itself. In domains such as human trafficking, marked by contested definitions, systemic injustice, and deep stakeholder asymmetries, such assumptions can obscure foundational ethical concerns. RQ offers an upstream, deliberative process for surfacing these concerns before design begins. Drawing from critical theory, participatory ethics, and relational responsibility, RQ formalizes a five-step framework to interrogate problem framings, confront techno-solutionist tendencies, and reflect on the moral legitimacy of intervention. Developed through interdisciplinary collaboration and engagement with survivor-led organizations, RQ was piloted in the domain of human trafficking (HT) which is a particularly high-stakes and ethically entangled application area. Its use led to a fundamental design shift: away from automated detection tools and toward survivor-controlled, empowerment-based technologies. We argue that RQ's novelty lies in both its temporal position, i.e, prior to technical design, and its orientation toward domains where harm is structural and ethical clarity cannot be achieved through one-size-fits-all solutions. RQ thus addresses a critical gap between abstract principles of responsible AI and the lived ethical demands of real-world deployment.

Introduction

Many "AI for good" initiatives operate under the assumption that technical products can adequately address complex social problems (Green 2019). Such efforts are frequently underpinned by a form of techno-solutionism (Heilinger 2022; Metcalf, Moss et al. 2019) which is the belief that social problems can be framed, simplified, and ultimately resolved through data-driven tools. This orientation simplifies multifaceted social challenges into narrowly defined problems and reduces systemic issues to tractable computational tasks, obscuring deeper structural, cultural, and political dynamics. In high-stakes domains such as human trafficking (HT), such oversimplification is not only misleading but dangerous: it can normalize surveillance-heavy, punitive approaches to social issues, marginalize survivor voices, and produce unintended harms under the guise of benevolent intervention.

While principles-based AI ethics frameworks (e.g., transparency, fairness, accountability) have become common in both academic and policy contexts (Leslie 2019; Madaio et al. 2020), particularly in socially sensitive domains (Reamer 2023; Hallamaa and Kalliokoski 2022), they tend to focus on how to design AI systems responsibly after a solution has already been deemed necessary. These frameworks offer important guidance on implementation (Steen, Neef, and Schaap 2021), yet they are often applied as post hoc checklists (Costanza-Chock 2020), lacking the reflexive space to interrogate whether AI should be built in the first place. As a result, dismantling techno-solutionist biases becomes significantly more challenging and ethical deliberation remains downstream of core design decisions.

This problem is compounded in AI-for-good initiatives by a Manichean worldview that frames the domain as unambiguously "bad" and the role of AI as necessarily "good" (Green 2019). This binary framing legitimizes intervention without reckoning with its potential harms or the situated complexities of the domain. In the case of human trafficking, for instance, the drive to "solve" the problem with technology can lead to invasive surveillance, misclassification of consensual sex work, and the silencing of survivor-defined priorities.

Recognizing this, recent work has called for more critical, upstream forms of ethical engagement. For example, the Studying Up framework (Kawakami et al. 2024a) advocates for self-reflection among researchers, collaboration with interdisciplinary experts, and the willingness to decline projects that are misaligned with justice-oriented goals. The Situate AI Guidebook (Kawakami et al. 2024b), a practical instantiation of these principles, provides structured prompts for public agencies to assess whether AI interventions should be pursued at all. Similarly, Bellini et al.'s Systematization of Knowledge on Digital-Safety Research (Bellini et al. 2024) focuses on participatory, low-risk methodologies for working with at-risk communities emphasizing early engagement, data privacy, and reciprocity.

The Human Rights AI Impact (HRAI) Assessment framework (Ontario Human Rights Commission n.d.), developed by the Law Commission of Ontario, provides questions to evaluate the human rights risks of AI tools and offers mitigation strategies. Unlike HRAI, RQ is not prescriptive about the actions to be taken based on the answers to its questions. Instead, it serves as a tool for self-reflection, enabling teams to navigate ethical dilemmas and make decisions grounded in their values and the specific context of their work. While not designed specifically for AI-for-good contexts, these works highlight the need for reflexivity and stakeholder involvement. However, they often lack a formalized process to guide upstream, contextual moral reasoning in domains as deeply fraught as HT.

To address this gap, we introduce Radical Questioning (RQ) - a five step, pre-design ethical framework developed to help AI practitioners interrogate the legitimacy of intervention before technical work begins. RQ builds on and extends prior work by formalizing a structured, iterative approach for surfacing assumptions, engaging with power asymmetries, and challenging normative narratives embedded in AI-for-good projects. Developed through interdisciplinary collaboration and sustained engagement with survivor-led organizations in the HT space, RQ is designed not only to guide more responsible project scoping, but also to empower practitioners to say no and walk away from AI solutions when ethical justification cannot be ensured.

By situating RQ within broader discussions of refusal, participatory design, and upstream ethics, this paper contributes a unique methodology tailored to domains where harm is structural, solutions are morally contested, and AI's role is far from self-evident. We demonstrate its application through a case study in human trafficking, where RQ helped reorient our project away from detection-based technologies and toward survivor-controlled tools. We argue that this kind of radical, pre-emptive inquiry is essential for technical communities working in ethically complex domains and propose RQ to that end. The contributions of this paper are:

1) we introduce Radical Questioning (RQ)—a structured, upstream ethics framework that helps AI practitioners assess whether an intervention should proceed, not just how to implement it responsibly.

2) we demonstrate RQ's value in the ethically complex domain of human trafficking, where it helped reframe our project from detection-based technologies to survivor-centered, empowerment-focused tools.

3) 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.

Section 2 reviews relevant literature on AI ethics principles, techno-solutionism, and prior work in the human trafficking domain. Section 3 introduces the Radical Questioning framework through an in-depth case study of its application in HT. Section 4 outlines key challenges and limitations. We conclude in Section 5 with reflections and takeaways.

Background

Ethics principles Contemporary AI ethics discourse, especially in AI-for-good contexts, is largely shaped by principles-based frameworks, fairness, accountability, transparency, privacy, and similar values, intended to apply across domains and jurisdictions (Leslie 2019; Commission et al. 2019; Fetic et al. 2020; Madaio et al. 2020). While widely adopted, these principles are often vague and decontextualized, offering limited guidance as developers move from abstract theory to concrete practice (Bleher and Braun 2023; Steen, Neef, and Schaap 2021). Institutional reviews show that ethical guidance is frequently shaped by private sector and governmental interests alike, often with instrumentalist aims, such as easing market deployment, satisfying compliance, or reducing litigation risk, rather than fostering deep ethical reflection (Hickok 2021; Jobin, Ienca, and Vayena 2019; Fjeld et al. 2020; Metcalf, Moss et al. 2019).

As Metcalf et al. (Metcalf, Moss et al. 2019) note, this "institutionalization of ethics" in tech companies tends to affirm corporate logic rather than critically interrogate them. As a result, principles-based approaches rarely enable the kind of radical self-questioning needed to confront techno-solutionist assumptions. When ethics is reduced to a checklist, developers can assume that building the tool is already justified, delaying deeper ethical engagement until post-hoc risk mitigation or compliance stages.

Responsible AI frameworks Even recent efforts to expand responsible AI (RAI) frameworks often fall short in addressing the contextual and relational nature of AI for social problems. The Situate AI Guidebook (Kawakami et al. 2024b), for example, offers a structured set of questions to help public agencies deliberate on AI project feasibility, operationalizing principles from frameworks like Studying Up (Kawakami et al. 2024a). These works call for early reflexivity, interdisciplinary collaboration, and the possibility of refusal principles closely aligned with RQ. However, RQ differs in offering a generalizable but flexible process for initiating this kind of ethical inquiry, particularly in contested domains like human trafficking, where problem framings and stakeholder legitimacy are deeply unstable.

Other approaches, such as Design Justice (Costanza-Chock 2020), emphasize participatory, community-led design processes. RQ can complement these efforts by critically interrogating the assumptions underlying a project's very existence ensuring that participatory efforts are not premised on flawed problem definitions. Similarly, Bellini et al.'s work on digital-safety research (Bellini et al. 2024) advocates low-risk methods and participatory safeguards, though it is more focused on qualitative research contexts than tool-building. Table 1 highlights the relation and differences between RQ and related ethics frameworks.

In light of these limitations, there is increasing advocacy for ethics frameworks that move beyond principle compliance and toward relational, human-centered, and reflexive approaches (Bleher and Braun 2023; Razi et al. 2021; Branford 2023; Zigon 2019; Heilinger 2022). Our contribution aligns with this relational turn. Rather than rejecting principles entirely, we argue that they must be complemented

by actionable, context-sensitive methods like Radical Questioning frameworks that help developers pause, build trust with affected communities, and take moral responsibility for whether to build at all.

Human Trafficking Human trafficking (HT), as defined in the Canadian Criminal Code, involves actions such as recruiting, transporting, or controlling individuals for the purpose of exploitation, most often sexual. While the law allows for both labor and sexual exploitation, in practice, HT legislation is increasingly used to regulate and criminalize sex work, often conflating consensual sex work with trafficking (House of Commons of Canada 2014; Durisin and van der Meulen 2020). The definition aligns with the Palermo Protocol (CRIME 2000), framing exploitation broadly and sometimes ambiguously.

Many AI tools in this domain rely on the assumption that traffickers post escort ads on behalf of victims treating these ads as reliable data sources for detection models (Thorn 2015; Crotty and Bouché 2018). While some survivors report having little input in how they are advertised, research also shows that agency can vary, with some sex workers influencing ad content or dynamics with third parties (Savoie-Gargiso and Morselli 2013). Nonetheless, detection-focused AI systems often mine adult services websites (ASWs) to flag suspected trafficking activity (Lee et al. 2021; Nair et al. 2022).

These tools are typically justified from a law enforcement perspective, aiming to speed up investigations or identify "risk signals." However, such interventions are not universally viewed as beneficial sex workers, survivors, and victims themselves often raise concerns about privacy, misidentification, and harm from carceral responses (Islam 2024). In this paper, we examine a representative ad-monitoring tool as a case study for applying Radical Questioning.

AI for combating HT Numerous AI initiatives target online escort ads, leveraging techniques such as knowledge graphs (Szekely et al. 2015), text processing and semi-supervised models (Dubrawski et al. 2015; Alvari, Shakarian, and Snyder 2017; Tong et al. 2017), entity extraction (Nagpal et al. 2017; Li et al. 2022; Liu et al. 2023), authorship attribution with financial linkage (Portnoff et al. 2017), image analysis (Stylianou, Souvenir, and Pless 2019), and multi-modal weak supervision (Nair et al. 2024). While technically sophisticated, these approaches often neglect ethical concerns and the lived experiences of trafficking survivors. Many focus on optimizing performance metrics using models like random forests (Dubrawski et al. 2015) and ensembles (Li et al. 2023; Mensikova and Mattmann 2018), without addressing data biases or validating impact on vulnerable populations.

Human-computer interaction (HCI) applications, including support chatbots (Maeng and Lee 2021) and survivor-focused apps (Gautam, Tatar, and Harrison 2020), prioritize usability but rarely engage deeply with AI ethics or potential harms. Similarly, visual analytics tools for law enforcement (Vajiac et al. 2022, 2023; Nair et al. 2022) may help identify trafficking patterns but risk reinforcing power imbalances. Demand-reduction strategies such as fake ads (Lugo-Graulich 2024) and behavioral profiling, often escape ethical scrutiny altogether.

Despite these efforts, many projects lack participatory design and risk retraumatization, privacy violations, and accountability gaps. More inclusive models, as proposed in (Feffer et al. 2023; Smith 2018; Witkin and Robjant 2018), advocate survivor-led, trauma-informed approaches. In our work, stakeholder engagement led us to abandon initial detection goals in favor of a survivor-controlled evidence tool, and to establish a survivor-led advisory board to guide its ethical use.

Ethics of anti-HT efforts A growing body of scholarship underscores the need to evaluate AI applications in human trafficking (HT) through ethical, participatory, and justice-oriented lenses. Deeb et al.(Deeb-Swihart, Endert, and Bruckman 2022) call for moving beyond principles-based frameworks to foster deeper ethical reflection in AI-for-HT research. The Design Justice framework by Costanza-Chock(Costanza-Chock 2020) similarly advocates for community-led processes that foreground marginalized voices and challenge the assumptions behind tool development. The RED Method (Rapid Ethical Deliberation)(Steen, Neef, and Schaap 2021) offers a practical complement by structuring stakeholder workshops to surface and address ethical concerns. Greene et al.(Greene, Hoffmann, and Stark 2019) question whether AI systems should be built at all if their normative goals fail to support equitable human flourishing, while Heilinger et al.(Heilinger 2022) call for a meta-ethical perspective that interrogates the foundational premises of AI ethics. Hickok(Hickok 2021) emphasizes the need for stronger accountability mechanisms and broader stakeholder inclusion in ethical deliberations.

Within the HT context specifically, several studies stress the importance of engaging affected communities in the design and evaluation of AI tools. Bhalerao et al.(Bhalerao 2022) highlight misalignments between stakeholder goals and caution against harmful outreach practices. Razi et al.(Razi et al. 2021) advocate for human-centered evaluations in online sexual risk detection, grounded in social and contextual nuance. Musto et al.(Musto and Boyd 2014) and Milivojevic(Milivojevic, Moore, and Segrave 2020) critique overly simplistic technological fixes that ignore broader structural dynamics. Chen et al. (Chen, Dell, and Roesner 2019) examine the trade-offs victim service providers face when balancing technological safety with client trust.

Collectively, these works point to critical gaps in current AI-for-HT approaches and advocate for more participatory, justice-driven, and context-sensitive alternatives. Building on these insights, the Radical Questioning (RQ) framework seeks to deepen ethical inquiry, foreground refusal and legitimacy, and center the lived experiences of marginalized groups throughout the design process.

Radical Questioning 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

Framework	Key Focus	Temporal Position	Approach	Relation to RQ
Principles-Based Ethics	High-level norms (e.g., fairness, transparency, privacy)	Post-design or during	Abstract, compliance-oriented	RQ complements by intervening before design; focuses on legitimacy, not just implementation.
Studying Up	Critique of power, interdisciplinary reflection, enabling refusal	Early design and policy	Reflexive, critical-theoretical	RQ operationalizes its ethos via a structured, domain-sensitive process.
Situate AI Guidebook	Structured deliberation for public-sector AI feasibility	Pre-design	Question-based, practitioner-guided	Shares intent with RQ but RQ offers a more generalizable and flexible structure.
Design Justice	Community-led design, equity, redistribution of power	Throughout design	Participatory, justice-oriented	RQ helps ensure problem framings are just and community-aligned from the outset.
Digital-Safety Research SoK	Safer practices for working with at-risk communities	During research design	Risk-aware, participatory	RQ is complementary but focused on tool development rather than research protocols.
Rapid Ethical Deliberation (RED)	Risk-focused deliberation on specific AI applications	Mid-design	Procedural, domain-specific	RQ differs by engaging before design and focusing on moral legitimacy.

Table 1: Comparison of Radical Questioning (RQ) with related frameworks

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 framework emerged from sustained engagement with survivor-led organizations and interdisciplinary collaborators, and was deeply informed by the legal and social tensions unique to the HT domain. While rooted in this context, RQ's five-step process is generalizable to other ethically complex domains, provided the questions it raises are grounded in domain-specific realities. RQ is intended for AI developers, researchers, policymakers, and interdisciplinary teams engaged in socially impactful technology projects, particularly where stakes are high, definitions are contested, and harm is unevenly distributed. It serves as a guide not for implementation, but for evaluating whether a proposed AI tool should exist at all.

In our case study, we applied RQ to the development of an AI tool in the HT space. Our team included AI researchers, criminologists, legal scholars, ethics experts, and HT survivors. We began by identifying the core motivations and assumptions driving the initiative, initially aiming to "solve" trafficking through data-driven detection. Early reviews of technical literature in this space presented a law enforcement-centric view, emphasizing pattern recognition and automation (Nair et al. 2024; Vajiac et al. 2022, 2023; Lee et al. 2021; Nair et al. 2022). However, as we engaged deeply with stakeholders, including survivor-led advocacy groups and frontline practitioners, we uncovered alternative narratives, ethical tensions, and the risks of reinforcing carceral harm. These insights led us to fundamentally rethink the project: shifting from detection to a survivor-empowerment tool focused on controlled documentation and consent. At every stage, we asked radical questions to uncover assumptions, redefine objectives, and assess legitimacy. We describe this process through the five steps of the RQ framework where each step is accompanied by example questions. These questions are illustrative of the types of inquiries the framework prompts, but in our case, adapted and refined based on the specific context of HT. Gray boxes throughout the next section include actual questions our team engaged with during the design process. We also present RQ diagrammatically in Figure 1.

Step 1: Defining 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 found that much of the literature and tooling assumes a fixed, binary notion of exploitation.

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 (Konrad et al. 2023; Brown 2024). These legal framings inform and reinforce AI problem definitions that treat all online sex work as inherently suspicious. A common assumption in the literature and tool development is that ads on adult service websites are authored by traffickers or pimps on behalf of victims (Thorn 2015; Crotty and Bouché 2018; Bouché and Wittmer 2015). However, this view neglects the complexity of agency and power dynamics in sex work. Research and survivor testimony show that in many cases, individuals involved in the sex trade may collaboratively negotiate how

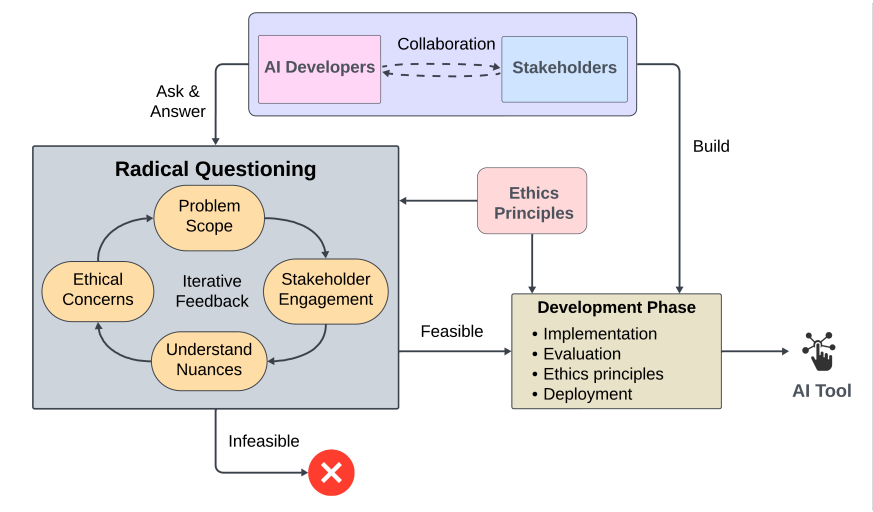

Figure 1: The proposed RQ framework is based on asking and answering radical questions through deliberative communication and collaboration between AI developers and stakeholders before development. This involves an iterative process of defining the problem scope, identifying affected communities, understanding nuances of the problem and mapping ethical concerns in order to determine the feasibility of an AI tool from an ethical perspective. If deemed ethically feasible, the tool is developed and if not, it is terminated.

their services are represented or even author the ads themselves (Savoie-Gargiso and Morselli 2013). Oversimplifying this relationship erases the agency of sex workers, reinforces carceral assumptions, and legitimizes surveillance-based interventions that can put marginalized communities at further risk. By unpacking these assumptions during this step, we recognized how initial framings can invisibilize lived experiences and distort the very problem AI systems aim to solve.

- What is human trafficking and why is it problematic? What is the social good pursued behind implementing such a tool that we think we can achieve?
- 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 we following technical feasibility without considering societal nuances and why?
- Is there a solution to the problem? Which are the solutions proposed for the problems? Who has constructed these "solutions"?
- Are there particular problems within the HT domain that we can focus on? What sort of resources are required for solving these problems?

Step 2: Identifying stakeholders and perspectives

Counter-HT efforts are highly fragmented, with stakeholders operating across different disciplines and contexts, often with minimal collaboration (see Figure 2). Each group tends to follow its own logic and priorities, leading to dichotomized perspectives, such as survivor-led versus non-survivor-led, empowering versus saving, or empirical versus ideological (Canada 2019). Effectively identifying and understanding these perspectives requires sustained engagement with diverse stakeholders, which demands significant resources in terms of time, funding, and workforce. Without adequate resources to ensure meaningful engagement, it may be better to halt the project than to proceed without this critical groundwork. Some of the main perspectives in HT are discussed below.

Governments At the highest-level, governmental organizations focus on public safety by introducing policies and strategies for fighting HT with technology (Lefebvre and Benyekhlef 2023). For example, in Canada, the government ratified the Palermo Protocol in 2002 to "prevent, suppress, punish trafficking in persons". The National Strategy to Combat Human Trafficking (2019-2024) increasingly emphasizes leveraging technology to address HT. Initiatives include hackathons, public-private collaborations, partnerships with tech companies and investments in technologies for AI detection. Law enforcement in the US and Canada also rely heavily on algorithms and tools involving image analyses and linking evidences and attributes in escort advertisements (Tong et al. 2017; Enrile and Aquino-Adriatico 2024) for intervention and charging of traffickers.

Human Rights Advocates In the second sphere, we have the human rights advocates, such as sex worker advocacy groups, and HT non-governmental organizations (including national trafficking hotlines) although the two have radically different viewpoints and collaborators. The former respect sex workers' agency and advocate for more safety measures for them. The latter fights more strictly against the exploitation of the vulnerable due to sexual merchandising. Academic research often critiques the police use of

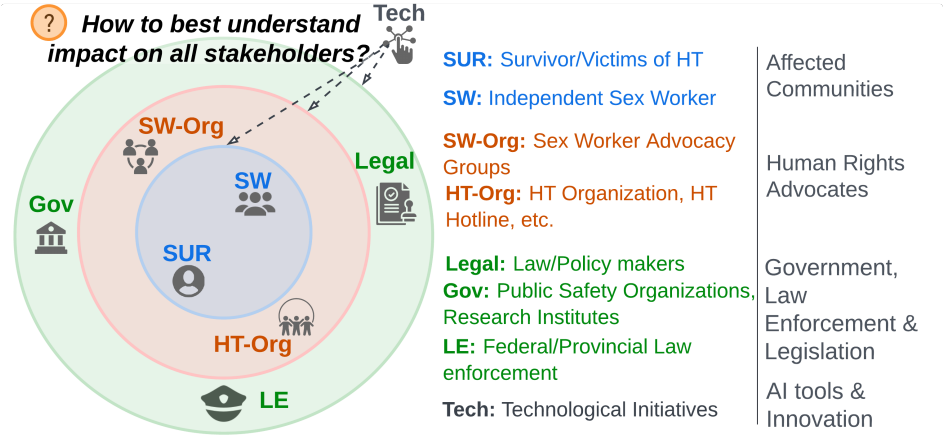

Figure 2: Non-exhaustive list of stakeholders within the HT domain, roughly categorized into three coaxial spheres. For example, survivors and sex workers are the communities directly affected by any technological initiatives involving surveilling online markets. Sex worker advocacy groups and organizations dedicated to supporting HT victims/survivors represent their respective interests and advocate for different understandings and methods to enhance human rights. Government focus on public safety and Parliament (legislative power) focus on fighting crime upholding minimal constitutional guarantees (such as protecting citizens' privacy.) Technological initiatives need to position themselves within this HT stakeholder ecosystem to be able to better understand its impacts (both positive and negative).

technology for countering HT from a human rights perspective (Deeb-Swihart, Endert, and Bruckman 2022; Milivojevic, Moore, and Segrave 2020) due to serious lack of transparency and collaboration between the police, government and human rights organizations (Rodrigues 2020; Lefebvre and Benyekhlef 2023).Their analytical framework is also limited to primarily analyzing the problem through a legal perspective. Some questions and problems, even if not legally relevant can still be relevant from a social, political or moral point of view.

Affected communities The innermost sphere depicts communities affected directly by the use of technological initiatives in the HT domain. Independent sex workers can be negatively impacted by the over-surveillance of online advertisements resulting in loss of work, living in constant fear, and reduced freedom of speech (called chilling effect). The point of view of affected communities are often mediated by sex workers' and survivors advocacy groups for various reasons including the difficulty of directly communicating with survivors themselves. This point of view can also be guarded due to distrust in the government, police, technological initiatives caused by historical discrimination and repression, and miscomprehension of the legislative framework (Farrell et al. 2019; Sterling and van der Meulen 2018a).

AI Practitioners The focus of most technological initiatives is typically on innovation and effectiveness. Any new tool for combating HT needs to be discussed with several, if not all, stakeholders to identify its best position within the HT ecosystem. When we consider the current AI initiatives in counter-HT, there are few works that discuss engagement with multiple stakeholders and/or study the impact of their technology on them (Maeng and Lee 2021; Gautam, Tatar, and Harrison 2020).

In our case study, we engaged a wide range of stakeholders, including law enforcement, public safety officials, survivor-led advocacy groups, human rights NGOs, and legal scholars. These groups brought divergent priorities while government actors focused on detection and prosecution, survivor advocates emphasized harm reduction, autonomy, and systemic change. Questions such as "Whose voice is missing?" and "Whose safety is prioritized, and at what cost?" surfaced tensions between punitive and restorative approaches. Recognizing that survivors are directly affected by counter-trafficking technologies, we shifted our focus from detection to co-designing survivor-centered tools grounded in trust, consent, and empowerment.

Building trust with survivor-led organizations required sustained, careful outreach. We identified groups led by those with lived experience and initiated contact with clear communication of our goals and openness to revising or halting the project. Rather than requesting input immediately, we asked how they might want to be involved. In many cases, it took multiple conversations to build rapport. Resources like the Trauma-Informed Code of Conduct (Witkin and Robjant 2018) and the Human Trafficking Survivor Leadership framework (Smith 2018) were critical to our approach. These relationships continue to shape our project, and we plan to conduct structured survivor surveys to generate actionable insights for future work.

- Who will be the end user of our tool? Who will own the tool and ensure its proper use? How will the incentives of the tool change depending on the owner?
- How do the different stakeholders implicated in the project understand the function and limits of the crimi-

nal law/and technology?

- How do law enforcement, policymakers, and advocacy groups differently perceive the scope of human trafficking, and how do these perceptions conflict or align?
- 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?

Step 3: Understanding nuances and complexities

By engaging with various stakeholders, we learned of several hyper-complexities of the HT domain which need to be taken into consideration before building an AI tool. We discuss some of them below.

Contesting notions of exploitation The interpretation of exploitation varies widely and is deeply influenced by social, cultural, and legal contexts. For instance, in Canada, law enforcement has used the notion of HT to target migrant sex workers and label groups supporting them as traffickers(Roots 2022). Policymakers and MPs increasingly conflate sex work with exploitation, leading to a broader, more ambiguous application of the term(Sterling and van der Meulen 2018b). In the literature of anti-HT AI initiatives, definitions of exploitation often rely on project-specific risk factors, sometimes intermixing consensual legal activities like "BDSM" with potentially exploitative practices (Giommoni and Ikwu 2021). Courts have been called upon to clarify this contested notion, which has even been challenged constitutionally for vagueness (sin 2020; uri 2013; gal 2019; dso 2016). Yet, the legal interpretations have tended to expand the concept, incorporating its diverse and subtle manifestations. When definitions of exploitation vary across legal, cultural, and societal contexts, the labels and risk factors chosen as ground truth can be inconsistent, biased, or overly simplistic. This variability propagates through the development of AI models, leading to tools that may misclassify consensual or lawful activities as exploitative or fail to identify nuanced cases of trafficking, ultimately reducing their reliability and ethical alignment.

Savior Complex The presumption that every individual experiencing trafficking wants or needs to be "saved", especially through law enforcement intervention, can be harmful (Heynen and van der Meulen 2022). Our consultations with survivors and survivor-led organizations, such as Women at the Center (Women at the Centre n.d.), reveal that survivors' needs and perceptions of justice are diverse and deeply influenced by personal circumstances, trauma, and societal factors. Some survivors may not feel ready or safe to press charges against their traffickers, while others may not fully comprehend the extent of their exploitation due to dependency, trauma, or differing understandings of what constitutes exploitation. Many face systemic barriers, such as fear of retaliation, lack of trust in judicial systems, or precarious immigration status, which complicate their paths to justice. Justice itself means different things to different survivors some prioritize pursuing education or securing residency over prosecuting their traffickers. Addressing HT requires moving beyond a narrow, savior-focused lens and adopting nuanced, survivor-centered approaches that address intersecting vulnerabilities and empower survivors in ways that respect their agency and diverse needs (Gerassi and Nichols 2021; Cha 2018).

Chilling effect While designing a tool for countering HT, there is a need to anticipate its harmful consequences and also realize that actions carried out with the aim of benefiting victims could in reality result in harming them and/or other members of affected communities, such as independent sex workers. AI tools that analyze the ad content looking for language cues or keywords can contribute to auto-censorship of sex workers and trafficked victims. For example, tools that flags certain keywords in the ads as suspicious may force independent sex workers to avoid them in their ads. This can create a climate of suspicion and fear of prosecution, resulting in auto-censure, known in legal and political literature as a "chilling effect" on free speech (Schauer 1978; Young 2023; Sterling and van der Meulen 2018a; Penney 2017). Hence, AI tools in this highly complex context creates an important dilemma - a tool designed to help victims may have side effects of constraining, impeding and impacting other potential victims and consensual sex workers in their work.

A deeper understanding of the hyper-complexity of this domain has deterred us from turning towards any sort of surveillance-based proactive detection of HT. Instead we focus on an adjacent problem that contributes to the proliferation and profitability of HT. Traffickers are at low risk of identification, prosecution and being sentenced due to lack of sufficient evidence for conviction. The justice system is over-reliant on the testimony of the victim which can lead to secondary victimization (Gershuni n.d.). Our consultations also taught us that oftentimes, the victims' testimony can be discounted due to stigma from the perception of lawyers and judges and/or inconsistencies due to trauma and drug use (Haskell and Randall 2019). This redirected us to design a tool that can help reduce the burden of proof on survivors and minimize their re-traumatization.

- 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?

Step 4: Mapping ethical concerns

Accountability AI accountability deals with deciding who should take responsibility for AI operations and algorithms should there be a breach of law (Balasubramaniam et al. 2023; Ada Lovelace Institute 2021). It attempts to distribute responsibility among the developers and users of an AI system (Deeb-Swihart, Endert, and Bruckman 2022), individually as well as at the organization level (Gutierrez, Marchant, and Michael 2021). Current interpretations of accountability focus more on legal accountability and less on the broader legitimacy of building such an AI tool in the first place, including its fairness or the degree of confidence that citizens, academics, experts, and impacted communities may have in it. This definition of accountability does not enable radical questioning on the legitimacy and unfair use of AI.

In our case study, we identified two main concepts of accountability in the HT domain. One puts emphasis on identifying perpetrators, as decided by the AI tool, that need to be taken accountable for facilitating HT (e.g pimps, traffickers, website hosts, etc) (Inter-Agency Coordination Group against Trafficking in Persons 2022). The second is the legal and ethical accountability of the AI tool's developers and users, principally in case of breach of privacy or misuse. One unique problem in this domain is that the act of holding others responsible for their contribution to HT and the accountability of the persons behind the AI initiative seem to be treated separately. We argue that the latter is inherently linked to the former and accountability needs to be seen as a dialogue between stakeholders and the AI developers with a formal structure for critiquing and challenging the decisions made with respect to the tool (Hulstijn 2023).

Privacy The privacy concerns discussed in current ethical guidelines focus more on the storage, collection, and access to data, the scope of data collection, its impact on marginalized groups, and security of datasets (Deeb-Swihart, Endert, and Bruckman 2022). AI researchers can tend to hire lawyers specialized in privacy law to assure that sensitive data are protected and to mitigate their risk of litigation. Privacy laws are usually country specific and tools built in one country may not be transferrable to others without radical changes in its design, use and aim. In Canada, private organisms cannot proactively and voluntarily collect, use and divulge information on the Internet without consent. The police need a warrant to indirectly access it through private organizations and are highly encouraged to obtain one to access personal information on the Internet (Robertson, Khoo, and Song 2020; Lefebvre and Benyekhlef 2023; Sanders et al. 2017). This can severely restrict the capacity of tools used by the police for analyzing data on the web to detect suspicious patterns.

Even if the privacy laws cannot offer clarity on complex problems and an AI tool may be in a gray zone, one might want to understand what privacy means for the persons affected by such a tool (Konrad et al. 2023). For example, in our case study, independent sex workers may not wish their privacy to be compromised as it may affect their livelihood. Sometimes victims do not wish their identities be revealed. Even if an AI tool does not use person-identifiable information, its outcomes can have subsequent impacts that compromises personal privacy. For example, tools that detect patterns of trafficking online are often imperfect and result in false positives. These can initiate follow up investigations resulting in the revelation of the individuals involved and adversely affecting them, especially voluntary sex workers. Even if they are not directly arrested, criminalized and/or prosecuted on the basis of the algorithm detection, increasing their visibility to the eyes of law enforcement can be detrimental for these persons and communities. As we engaged more with concrete groups affected by the tool, it became clearer that intrusion of privacy can result in precise and concrete harms. Thus, the viability of an AI tool goes beyond referencing the main ethical principles and privacy laws. Moreover, privacy law frameworks have not necessarily been designed to accommodate scenarios where private initiatives implement a powerful technology like AI to help the police with surveillance. Lawyers and privacy law specialists can help mitigate the risk of a breach of privacy but cannot answer many vital questions like what privacy means for the individuals affected. This brings the responsibility back to the AI developers to frame complex questions, identify problems and solve dilemmas themselves.

Fairness and Equality In the context of AI tools, equality, although initially seen through the lens of biased training data and reducing false positives, has been redefined to include more complex and pluralistic perspectives (Long 2021; Jain et al. 2024). We focus on the relationship between equality and true positives. The true positives predicted by an AI algorithm are inherently dependent on the ground-truth examples used to train it. In the HT context, these ground-truth are examples that are manually-labeled as "exploitation". However, this notion of exploitation is not a scientifically demonstrable and invariable reality. Our understanding of exploitation can be tainted by biases around sexuality, race, age, sex and gender (Musto and Boyd 2014; Sanders 2018; Giommoni and Ikwu 2021; Wijers 2015). What are the conditions that render a relationship between two individuals as exploitative? Many different answers can be found in social sciences and law, as the notion of exploitation has evolved with time and depends on the context. An analysis of electronic transcripts of communication between sex workers and their pimp in a criminal network in Canada, has revealed how the sex workers have non-negligible control over their pimp (Morselli and Savoie-Gargiso 2014). Some studies also show that sex workers can also play the role of recruiter and facilitator (Wijkman and Kleemans 2019; Konrad et al. 2023). Thus, the frontier between exploiter and exploited may not be as clear-cut as one thinks and it is important to question the relative ground truth on which the algorithms are trained.

Several methods also rely on risk factors and suspicious keywords for detecting possible exploitation (Kennedy 2012; Zhu, Li, and Jones 2019; Dubrawski et al. 2015; Tong et al. 2017; Nair et al. 2024) which, again, are not based on an immutable and objective reality. These premises, and the bias and prejudices that sustain them, can result in the

unequal over-surveillance and over-criminalization of certain persons and groups. For example, keywords describing ethnicities such as "Asian" and "Eastern European" may become the focus as past investigations of police in Canada have mostly targeted these populations due to their apparent vulnerable status to transnational traffic (Sanders 2018; Ibanez and Gazan 2016).

In addition, proxies such as "petite" or "small size" for detecting child sexual exploitation, can replicate over-focus on Asian sex workers. Similarly, flagging ads because of "poor language" or use of a "foreign language" can have similar effects. An AI tool that reproduces these assumptions, directly or indirectly, can be highly problematic when considered through the lens of equality. Thus, before discussing false and true positives in the context of equality, we need to ask how the ground-truth was built. This further illustrates previous research (Green 2019) stating that a technological project that hides itself under the "social good" framework and does not engage in a reflexive engagement with social and political context risks of reproducing and exacerbating the exact forms of social oppression that they try to dismantle.

Transparency Current understanding of transparency is primarily limited to revealing the nature of the training data used, the goal and scope of the tool, its limitations and how it arrives at conclusions (Gutierrez and Marchant 2021; Jobin, Ienca, and Vayena 2019). Transparency is often linked with public access to source code and documentation regarding the data, model and evaluation (Eiras et al. 2024). Many AI solutions are deemed transparent simply due to their open-source nature. Public policies conceive transparency as obtaining explanations of algorithmic systems, such that individuals can learn of their use and demand answers (Ada Lovelace Institute 2021). However, solely using open-source techniques is not enough.

The principle-form of transparency does not provide sufficient information on whether or not the AI researchers have seriously engaged with the complexity of the social problem they are attempting to solve. It does not disclose who they have worked with and the nature of their collaboration. It also does not allow questioning of the interests of the AI team involved and their incentives and motivations for building the tool. For these reasons, we argue that placing transparency into the principle framework obscures radical questions about what precedes the design and implementation of the tool.

To avoid this, previous works have suggested that computer scientists explicitly consider and articulate the commitments behind their work, demonstrate that they have tried to find less intrusive alternative solutions and explain why the contemplated solution is preferable (Green 2019). We echo this sentiment and declare the need for thorough cross-checking and verification of the algorithmic outputs by relevant stakeholders and communities impacted by the tool.

Moreover, several AI tools are evaluated using metrics and parameters that are not necessarily relevant in their respective domains and are not evaluated in the way they are meant to be used. Such metrics are not always explained in a manner understandable to lay persons, particularly those using and affected by the tool.

In our case study, the proposed survivor-centric tool responds to a clearly articulated need: enabling survivors to document and manage their own evidence of exploitation in a secure, controlled manner. Unlike detection systems designed for third-party surveillance, this tool is built around survivor autonomy, with the primary motivation being to support self-advocacy, justice-seeking, and recovery on the survivor's terms. Ethical concerns such as retraumatization, data misuse, and privacy violations were central to the design process. Because the tool is intended for voluntary use by survivors themselves, it allows for granular control over what is documented, when, and how. Fairness and accuracy will be evaluated through participatory user studies, pilot deployments, and continuous feedback from survivor-led organizations and advocacy groups. To ensure sustained accountability and legitimacy, we have convened an interdisciplinary advisory board, including survivors, survivor leaders and legal experts, that will oversee the tool's development, deployment, and evaluation.

- 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? Can such fairness metrics even be measured efficiently without first deploying the solutions or conducting user-studies?
- How to choose who needs to take accountability for the tool? What does accountability mean to those involved in the domain?
- Was the need for the tool expressed by the persons it surveils, tries to protect and/or rescue from "exploitation"?
- What according to the team and different stakeholders is considered the legitimate and appropriate use of the tool? What purposes can and cannot be served by the tool? By whom should it not be used?
- 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 behaviors can be seen as physical and psychological coercion for the provision of sexual services? Does money in exchange for sexual services necessarily vitiate consent? Is it inherently exploitative? Is a woman that works with a male "pimp" necessarily exploited? How can we account for these complexities in the tool?
- What are the incentives, interests and motivation for developing an AI tool that will integrate the HT ecosystem (both financial and structural)? What is the composition of the team behind the tool?

Step 5: Iterative Feedback

In our work, continuous engagement with affected communities, particularly survivors, was central to shaping an ethical and responsive design process. We established iterative

feedback mechanisms early on, relying on both structured and informal channels to gather input, while remaining attentive to the risks of retraumatization (Witkin and Robjant 2018). To protect survivor well-being, we prioritized low-pressure, trauma-informed modes of consultation, such as informal consultations and facilitated workshops. As feedback came in, we made a conscious effort to reflect on the full spectrum of responses, including those that challenged our initial goals or assumptions. This helped us recognize when we were at risk of selective listening or confirmation bias (Smith 2018). In some instances, survivor input led us to revise our design such as deciding what information should be displayed, reframe core functionalities or postpone decisions such as regarding how to verify users, until concerns could be better addressed. We also drew on resources like the Trauma-Informed Code of Conduct (Witkin and Robjant 2018) and expert-led trainings to better understand best practices for engaging with vulnerable communities. Throughout, we asked ourselves not just whether we had heard the feedback, but whether we had meaningfully acted on it and whether the project still aligned with the needs and well-being of those it aimed to support.

- 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?

Challenges and Limitations

While the Radical Questioning (RQ) framework proved valuable in our case study, it also presents several challenges and limitations. First, RQ is intentionally non-prescriptive. Its purpose is to surface ethical tensions and challenge foundational assumptions not to offer ready-made answers or technical implementation plans. This can be difficult in fast-paced, outcome-driven settings where teams are seeking concrete guidance.

Second, the effectiveness of RQ depends heavily on authentic stakeholder engagement. Building trust with affected communities, especially those marginalized or criminalized, such as sex workers and trafficking survivors, requires time, care, and institutional support. In our case, it involved sustained outreach and relationship-building with survivor-led organizations, often across multiple layers of communication and consent. Such access may not always be readily available and can be mediated by complex gatekeeping structures.

Third, RQ's impact is shaped by the positionality and reflexivity of the development team. It requires teams to be open to rethinking project goals, confronting discomfort, and taking personal and collective responsibility for ethical choices. Without this willingness, RQ risks becoming a symbolic gesture rather than a meaningful intervention.

Fourth, RQ's transferability is limited by context. While the five-step process is generalizable, the specific questions it prompts must be re-grounded in the histories, power dynamics, and lived realities of each domain. What counts as harm, justice, or legitimacy in one context may not apply in another. Some of the ethical dilemmas and questions that we have presented may change depending on 1) who might own the product, for example, a law office, a research institute, a private company or a public agency 2) the degree of automation of the tool – a completely autonomous decision-maker vs a human-in-the-loop solution and 3) the proposed use-case for the tool.

Finally, we acknowledge that maintaining trust is more important than simply adhering to abstract ethical principles. The viability of any tool developed through RQ depends not just on legal compliance, but on sustained relationships with the communities it affects. Ownership structures, degrees of automation, and use cases all influence the ethical stakes and must be considered as part of the inquiry. RQ is best viewed not as a checklist, but as a practice, one that demands time, humility, and ongoing ethical commitment.

Conclusions and Takeaways

Radical Questioning (RQ) offers a necessary shift in how we think about responsible AI: not as a checklist applied post-design, but as a deeper, pre-project ethics practice. Its purpose is not to prescribe solutions, but to create space for deliberation, inviting developers to examine their assumptions, consider alternate framings, and ask whether a system should be built at all. We hope this framework inspires AI practitioners to begin with questions, not code: to engage with affected communities early, remain transparent about ethical uncertainties, and embrace the possibility of walking away if ethical development is not possible.

In our work on human trafficking (HT), RQ fundamentally reshaped our trajectory. What began as a project aimed at detection and intervention evolved through survivor consultations and critical reflection into a tool centered on autonomy, consent, and care. We learned that justice in this domain is not always about prosecution or rescue, but about restoring agency and meeting immediate needs. This reorientation demonstrates the value of radical questioning in uncovering overlooked harms and aligning technical work with the lived realities of those most affected.

The broader challenge lies in the culture of AI development itself. Current incentive structures reward speed, novelty, and generalizability, often at the expense of ethical rigor and contextual understanding. To create technologies that are truly accountable and inclusive, AI research must move beyond surface-level ethics statements and integrate sustained, context-sensitive ethical inquiry into its core practices. For this to happen, the values embodied by frameworks like RQ must be taken seriously across both technical venues and ethics-focused communities. Only then can we begin to build AI that reflects not just what is possible, but what is just.

Acknowledgments

We are grateful to Ollie, Christos, Catalina, Mark and Kevin for their generous insights and constructive feedback, which meaningfully shaped the direction and clarity of this work.

References

2013. R. c. Urizar, 2013 QCCA 46. Available at CanLII: https://www.canlii.org/en/qc/qcca/doc/2013/2013qcca46/2013qcca46.html.

2016. R. v. D'Souza, 2016 ONSC 2749. Available at CanLII: https://www.canlii.org/en/on/onsc/doc/2016/2016onsc2749/2016onsc2749.html.

2019. R. v. Gallone, 2019 ONCA 663. Available at CanLII: https://www.canlii.org/en/on/onca/doc/2019/2019onca663/2019onca663.html.

2020. R. v. Sinclair, 2020 ONCA 61. Available at CanLII: https://www.canlii.org/en/on/onca/doc/2020/2020onca61/2020onca61.html.

Ada Lovelace Institute, O., AI Now. 2021. Algorithmic accountability for the public sector. https://www.opengovpartnership.org/wp-content/uploads/2021/08/executive-summary-algorithmic-accountability.pdf.

Alvari, H.; Shakarian, P.; and Snyder, J. K. 2017. Semi-supervised learning for detecting human trafficking. *Security Informatics*, 6(1): 1.

Balasubramaniam, N.; Kauppinen, M.; Rannisto, A.; Hiekkanen, K.; and Kujala, S. 2023. Transparency and explainability of AI systems: From ethical guidelines to requirements. *Information and Software Technology*, 159: 107197.

Bellini, R.; Tseng, E.; Warford, N.; Daffalla, A.; Matthews, T.; Consolvo, S.; Woelfer, J. P.; Kelley, P. G.; Mazurek, M. L.; Cuomo, D.; et al. 2024. Sok: Safer digital-safety research involving at-risk users. In *2024 IEEE Symposium on Security and Privacy (SP)*, 635–654. IEEE.

Bhalerao, R. 2022. *Analyzing Harms of Online Platform and Policy Design*. New York University Tandon School of Engineering.

Bleher, H.; and Braun, M. 2023. Reflections on putting AI ethics into practice: how three AI ethics approaches conceptualize theory and practice. *Science and Engineering Ethics*, 29(3): 21.

Bouché, V.; and Wittmer, D. E. 2015. Gendered diffusion on gendered issues: the case of human trafficking. *Journal of Public Policy*, 35(1): 1–33.

Branford, J. 2023. Experiencing AI and the Relational 'Turn'in AI Ethics. In *International Conference on Computer Ethics*, volume 1.

Brown, S. 2024. Policing sex work online: sex workers' views on the risks and benefits of using AI to police online ads for sexual services.

Canada, P. S. 2019. National strategy to combat human trafficking 2019–2024. *Government of Canada*.

Cha, S. 2018. Deconstructing human trafficking and victimization: A reshaping of Canadian discourses and policies. *International Human Rights Internship Program: Working Paper Series*, 6(12): 1–49.

Chen, C.; Dell, N.; and Roesner, F. 2019. Computer security and privacy in the interactions between victim service providers and human trafficking survivors. In *28th USENIX Security Symposium (USENIX Security 19)*, 89–104.

Commission, E. H. A.-E.; et al. 2019. Independent High-Level Expert Group on Artificial Intelligence (2019). *Ethics guidelines for trustworthy AI*.

Costanza-Chock, S. 2020. *Design justice: Community-led practices to build the worlds we need*. The MIT Press.

CRIME, A. T. O. 2000. PROTOCOL TO PREVENT, SUPPRESS AND PUNISH TRAFFICKING IN PERSONS, ESPECIALLY WOMEN AND CHILDREN, SUPPLEMENTING THE UNITED NATIONS CONVENTION AGAINST TRANSNATIONALORGANIZED CRIME.

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

Deeb-Swihart, J.; Endert, A.; and Bruckman, A. 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): 1–29.

Dubrawski, A.; Miller, K.; Barnes, M.; Boecking, B.; and Kennedy, E. 2015. Leveraging publicly available data to discern patterns of human-trafficking activity. *Journal of Human Trafficking*, 1(1): 65–85.

Durisin, E. M.; and van der Meulen, E. 2020. Sexualized Nationalism and Federal Human Trafficking Consultations: Shifting Discourses on Sex Traff.

Eiras, F.; Petrov, A.; Vidgen, B.; Schroeder, C.; Pizzati, F.; Elkins, K.; Mukhopadhyay, S.; Bibi, A.; Purewal, A.; Botos, C.; et al. 2024. Risks and opportunities of open-source generative ai. *arXiv preprint arXiv:2405.08597*.

Enrile, A.; and Aquino-Adriatico, G. 2024. *Technology Innovations in Fighting Slavery and Human Trafficking*, 179–203. Cham: Springer Nature Switzerland. ISBN 978-3-031-58614-9.

Farrell, A.; Dank, M.; de Vries, I.; Kafafian, M.; Hughes, A.; and Lockwood, S. 2019. Failing victims? Challenges of the police response to human trafficking. *Criminology & Public Policy*, 18(3): 649–673.

Feffer, M.; Skirpan, M.; Lipton, Z.; and Heidari, H. 2023. From preference elicitation to participatory ML: A critical survey & guidelines for future research. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 38–48.

Fetic, L.; Fleischer, T.; Grünke, P.; Hagendorf, T.; Hallensleben, S.; Hauer, M.; Herrmann, M.; Hillerbrand, R.; Hustedt, C.; Hubig, C.; et al. 2020. From Principles to Practice. An interdisciplinary framework to operationalise AI ethics.

Fjeld, J.; Achten, N.; Hilligoss, H.; Nagy, A.; and Srikumar, M. 2020. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication*, (2020-1).

Gautam, A.; Tatar, D.; and Harrison, S. 2020. Crafting, communality, and computing: Building on existing strengths to support a vulnerable population. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.

Gerassi, L.; and Nichols, A. 2021. Social work education that addresses trafficking for sexual exploitation: An intersectional, anti-oppressive practice framework. *Anti-trafficking review*, (17): 20–37.

Gershuni, R. n.d. Quote on Evaluating Evidence in Human Trafficking Cases. International Legal Expert on Human Trafficking. "In view of the typical weaknesses that plague victim testimonies, it is necessary to gather other forms of evidence and evaluate on the totality of the evidence rather than limiting it to the victim statement.".

Giommoni, L.; and Ikwu, R. 2021. Identifying human trafficking indicators in the UK online sex market. *Trends in Organized Crime*, 1–24.

Green, B. 2019. Good" isn't good enough. In *Proceedings of the AI for Social Good workshop at NeurIPS*, volume 17.

Greene, D.; Hoffmann, A. L.; and Stark, L. 2019. Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning.

Gutierrez, C. I.; and Marchant, G. E. 2021. Soft Law 2.0: Incorporating incentives and implementation mechanisms into the governance of artificial intelligence. *Organisation for Economic Co-operation and Development. https://www. oecd. ai/wonk/soft-law-2-0.2021*.

Gutierrez, C. I.; Marchant, G. E.; and Michael, K. 2021. Effective and trustworthy implementation of AI soft law governance. *IEEE Transactions on Technology and Society*, 2(4): 168–170.

Hallamaa, J.; and Kalliokoski, T. 2022. AI ethics as applied ethics. *Frontiers in computer science*, 4: 776837.

Haskell, L.; and Randall, M. 2019. *The impact of trauma on adult sexual assault victims*. Justice Canada.

Heilinger, J.-C. 2022. The ethics of AI ethics. A constructive critique. *Philosophy & Technology*, 35(3): 61.

Heynen, R.; and van der Meulen, E. 2022. Anti-trafficking saviors: Celebrity, slavery, and branded activism. *Crime, media, culture*, 18(2): 301–323.

Hickok, M. 2021. Lessons learned from AI ethics principles for future actions. *AI and Ethics*, 1(1): 41–47.

House of Commons of Canada. 2014. Bill C-36, Protection of Communities and Exploited Persons Act: Second Reading, House of Commons Debates, 41st Parliament, 2nd Session, Sitting No. 101. Remarks by Hon. Peter MacKay at 5:10 p.m. and 5:25 p.m. (June 11, 2014).

Hulstijn, J. 2023. Computational accountability. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, 121–130.

Ibanez, M.; and Gazan, R. 2016. Virtual indicators of sex trafficking to identify potential victims in online advertisements. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 818–824. IEEE.

Inter-Agency Coordination Group against Trafficking in Persons. 2022. Use and abuse of technology: Joint Statement of the Inter-Agency Coordination Group against Trafficking in Persons (ICAT) on the World Day against Trafficking in Persons. Statement, ICAT.

Islam, F. 2024. Human Trafficking Law Enforcement Over the Victims and Offenders: The Perspective of Anti-Trafficking Stakeholders. *Victims & Offenders*, 1–29.

Jain, S.; Suriyakumar, V.; Creel, K.; and Wilson, A. 2024. Algorithmic Pluralism: A Structural Approach To Equal Opportunity. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 197–206.

Jobin, A.; Ienca, M.; and Vayena, E. 2019. The global landscape of AI ethics guidelines. *Nature machine intelligence*, 1(9): 389–399.

Kawakami, A.; Coston, A.; Zhu, H.; Heidari, H.; and Holstein, K. 2024a. The Situate AI Guidebook: Co-Designing a Toolkit to Support Multi-Stakeholder, Early-stage Deliberations Around Public Sector AI Proposals. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703300.

Kawakami, A.; Coston, A.; Zhu, H.; Heidari, H.; and Holstein, K. 2024b. The Situate AI Guidebook: Co-Designing a Toolkit to Support Multi-Stakeholder, Early-stage Deliberations Around Public Sector AI Proposals. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703300.

Kennedy, E. 2012. Predictive patterns of sex trafficking online. *Dietrich College Honors Theses*, 1–45.

Konrad, R. A.; Maass, K. L.; Dimas, G. L.; and Trapp, A. C. 2023. Perspectives on how to conduct responsible anti-human trafficking research in operations and analytics. *European Journal of Operational Research*, 309(1): 319–329.

Lee, M.-C.; Vajiac, C.; Kulshrestha, A.; Levy, S.; Park, N.; Jones, C.; Rabbany, R.; and Faloutsos, C. 2021. InfoShield: Generalizable information-theoretic human-trafficking detection. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, 1116–1127. IEEE.

Lefebvre, G.; and Benyekhlef, K. 2023. "Predictive policing in Canada" in "Artificial Intelligence and Administration of Criminal Justice". *International Review of Penal Law*, 94.

Leslie, D. 2019. Understanding artificial intelligence ethics and safety. *arXiv preprint arXiv:1906.05684*.

Li, R.; Tobey, M.; Mayorga, M. E.; Caltagirone, S.; and Özaltın, O. Y. 2023. Detecting human trafficking: Automated classification of online customer reviews of massage businesses. *Manufacturing & Service Operations Management*, 25(3): 1051–1065.

Li, Y.; Nair, P.; Pelrine, K.; and Rabbany, R. 2022. Extracting person names from user generated text: Named-entity recognition for combating human trafficking. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2854–2868.

Liu, J.; Yu, H.; Sujaya, V.; Nair, P.; Pelrine, K.; and Rabbany, R. 2023. SWEET-Weakly Supervised Person Name Extraction for Fighting Human Trafficking. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 3355–3367.

Long, R. 2021. Fairness in machine learning: Against false positive rate equality as a measure of fairness. *Journal of Moral Philosophy*, 19(1): 49–78.

Lugo-Graulich, K. 2024. Indicators of Sex Trafficking in Online Escort Ads, 7 US states, 2013-2020.

Madaio, M. A.; Stark, L.; Wortman Vaughan, J.; and Wallach, H. 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.

Maeng, W.; and Lee, J. 2021. Designing a chatbot for survivors of sexual violence: Exploratory study for hybrid approach combining rule-based chatbot and ml-based chatbot. In *Proceedings of the Asian CHI Symposium 2021*, 160–166.

Mensikova, A.; and Mattmann, C. A. 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.

Metcalf, J.; Moss, E.; et al. 2019. Owning ethics: Corporate logics, silicon valley, and the institutionalization of ethics. *Social Research: An International Quarterly*, 86(2): 449–476.

Milivojevic, S.; Moore, H.; and Segrave, M. 2020. Freeing the Modern Slaves, One Click at a Time: Theorising human trafficking, modern slavery, and technology. *Anti-trafficking review*, (14): 16–32.

Morselli, C.; and Savoie-Gargiso, I. 2014. Coercion, control, and cooperation in a prostitution ring. *The ANNALS of the American Academy of Political and Social Science*, 653(1): 247–265.

Musto, J. L.; and Boyd, D. 2014. The trafficking-technology nexus. *Social Politics*, 21(3): 461–483.

Nagpal, C.; Miller, K.; Boecking, B.; and Dubrawski, A. 2017. An Entity Resolution Approach to Isolate Instances of Human Trafficking Online. In *Proceedings of the 3rd Workshop on Noisy User-generated Text, NUT@EMNLP*, 77–84.

Nair, P.; Li, Y.; Vajiac, C.; Olligschlaeger, A.; Lee, M.-C.; Park, N.; Chau, D. H.; Faloutsos, C.; and Rabbany, R. 2022. Vispad: Visualization and pattern discovery for fighting human trafficking. In *Companion Proceedings of the Web Conference 2022*, 273–277.

Nair, P.; Liu, J.; Vajiac, C.; Olligschlaeger, A.; Chau, D. H.; Cazzolato, M.; Jones, C.; Faloutsos, C.; and Rabbany, R. 2024. T-NET: Weakly Supervised Graph Learning for Combatting Human Trafficking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 22276–22284.

Ontario Human Rights Commission. n.d. Human Rights and AI: Impact Assessment. Accessed: 2025-01-21.

Penney, J. 2017. Internet surveillance, regulation, and chilling effects online: A comparative case study. *Regulation, and Chilling Effects Online: A Comparative Case Study (May 27, 2017)*, 6(2).

Portnoff, R. S.; Huang, D. Y.; Doerfler, P.; Afroz, S.; and McCoy, D. 2017. Backpage and Bitcoin: Uncovering Human Traffickers. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Razi, A.; Kim, S.; Alsoubai, A.; Stringhini, G.; Solorio, T.; De Choudhury, M.; and Wisniewski, P. J. 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): 1–38.

Reamer, F. G. 2023. Artificial intelligence in social work: Emerging ethical issues. *International Journal of Social Work Values and Ethics*, 20(2): 52–71.

Robertson, K.; Khoo, C.; and Song, Y. 2020. To surveil and predict: A human rights analysis of algorithmic policing in Canada.

Rodrigues, R. 2020. Legal and human rights issues of AI: Gaps, challenges and vulnerabilities. *Journal of Responsible Technology*, 4: 100005.

Roots, K. 2022. *The domestication of human trafficking: Law, policing, and prosecution in Canada*. University of Toronto Press.

Sanders, T. 2018. Enhancing the study of sex work. *Sexualities*, 21(8): 1346–1350.

Sanders, T.; Scoular, J.; Campbell, R.; Pitcher, J.; and Cunningham, S. 2017. *Internet sex work: Beyond the gaze*. Springer.

Savoie-Gargiso, I.; and Morselli, C. 2013. Homme à femmes: le proxénète et sa place parmi les prostituées. *Criminologie*, 46(1): 243–268.

Schauer, F. 1978. Fear, risk and the first amendment: Unraveling the chilling effect. *BUL rev.*, 58: 685.

Smith, M. 2018. Human trafficking survivor leadership in the United States. *Freedom Network USA. Retrieved October*, 19(2020): 2019–1.

Steen, M.; Neef, M.; and Schaap, T. 2021. A method for rapid ethical deliberation in research and innovation projects. *International Journal of Technoethics (IJT)*, 12(2): 72–85.

Sterling, A.; and van der Meulen, E. 2018a. "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): 291–308.

Sterling, A.; and van der Meulen, E. 2018b. "We Are Not Criminals": Sex Work Clients in Canada and the Constitution of Risk Knowledge. *Canadian Journal of Law and Society / Revue Canadienne Droit et Société*, 33(3): 291–308.

Stylianou, A.; Souvenir, R.; and Pless, R. 2019. TraffickCam: Explainable Image Matching For Sex Trafficking Investigations. *arXiv preprint arXiv:1910.03455*.

Szekely, P.; Knoblock, C. A.; Slepicka, J.; Philpot, A.; Singh, A.; Yin, C.; Kapoor, D.; Natarajan, P.; Marcu, D.; Knight, K.; et al. 2015. Building and using a knowledge graph to combat human trafficking. In *International Semantic Web Conference*, 205–221. Springer.

Thorn, D. V. B. 2015. A Report on the Use of Technology to Recruit, Groom and Sell Domenstic Minor Sex Trafficking Victims. https://www.thorn.org/wp-content/uploads/2015/02/Survivor_Survey_r5.pdf. Accessed: 2024-02-09.

Tong, E.; Zadeh, A.; Jones, C.; and Morency, L.-P. 2017. Combating human trafficking with deep multimodal models. *arXiv preprint arXiv:1705.02735*.

Vajiac, C.; Chau, D. H.; Olligschlaeger, A.; Mackenzie, R.; Nair, P.; Lee, M.-C.; Li, Y.; Park, N.; Rabbany, R.; and Faloutsos, C. 2022. TRAFFICVIS: visualizing organized activity and spatio-temporal patterns for detecting and labeling human trafficking. *IEEE transactions on visualization and computer graphics*, 29(1): 53–62.

Vajiac, C.; Chau, D. H.; Olligschlaeger, A.; Nair, P.; Lee, M.-C.; Cazzolato, M. T.; Rabbany, R.; Faloutsos, C.; and Jones, C. 2023. TRAFFICBOARD: Digital Spatio-Temporal Pinboard for Human Trafficking Detection.

Wijers, M. 2015. Purity, victimhood and agency: Fifteen years of the UN trafficking protocol. *Anti-Trafficking Review*, (4).

Wijkman, M.; and Kleemans, E. 2019. Female offenders of human trafficking and sexual exploitation. *Crime, Law and Social Change*, 72: 53–72.

Witkin, R.; and Robjant, K. 2018. The Trauma-Informed Code of Conduct. *London: Helen Bamber Foundation*.

Women at the Centre. n.d. Women at the Centre. Accessed: 2025-01-21.

Young, H. 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): 125–139.

Zhu, J.; Li, L.; and Jones, C. 2019. Identification and detection of human trafficking using language models. In *2019 European Intelligence and Security Informatics Conference (EISIC)*, 24–31. IEEE.

Zigon, J. 2019. Can machines be ethical? On the necessity of relational ethics and empathic attunement for data-centric technologies. *Social Research: An International Quarterly*, 86(4): 1001–1022.