A Responsible Research and Innovation Approach to Assessing the Implications of Anti-Trafficking Technologies

TOWERA MOYO, Department of Computer Science, University of Oxford, Oxford, United Kingdom of Great Britain and Northern Ireland
WAEL ALBAYAYDH, Department of Computer Science, University of Oxford, Oxford, United Kingdom of Great Britain and Northern Ireland
HELENA WEBB, School of Computer Science, University of Nottingham, Nottingham, United Kingdom of Great Britain and Northern Ireland
OMER GUNES, Department of Computer Science, University of Oxford, Oxford, United Kingdom of Great Britain and Northern Ireland
MARINA JIROTKA, Department of Computer Science, University of Oxford, Oxford, United Kingdom of Great Britain and Northern Ireland

Responsible Research and Innovation (RRI) is crucial to address the risks associated with emerging technologies with the aim of minimising negative consequences and promoting positive outcomes. This article uses RRI principles and semi-structured interviews to examine the implications of tools designed to combat human trafficking. Although traffickers use technology to recruit and exploit victims, it also equips stakeholders, such as governments, policymakers, researchers, and anti-trafficking experts, to combat these crimes. This article focuses on the impact of anti-trafficking software systems that identify victims and human traffickers in the United States and the United Kingdom, highlighting the positive and negative consequences of such tools and proposing strategies to mitigate unintended negative consequences, thereby promoting the development of responsible anti-trafficking tools. In addition, this article demonstrates the practical implementation of RRI within the information and communication technology sector, with a specific focus on the evaluation of existing software systems.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; **User studies;;** • **Security and privacy** → Social aspects of security and privacy; • **Social and professional topics** → *Socio-technical systems*; • **General and reference** → **Empirical studies**; *Evaluation*; **Design**; *Reliability*; *Performance*;

Additional Key Words and Phrases: Responsible research and innovation, human trafficking, anti-trafficking technologies, privacy, ethics, societal impact, software systems, ethical considerations, interviews, qualitative research, design considerations

Authors' Contact Information: Towera Moyo, Department of Computer Science, University of Oxford, Oxford, Oxfordshire, United Kingdom of Great Britain and Northern Ireland; e-mail: towera.moyo@yahoo.com; Wael Albayaydh, Department of Computer Science, University of Oxford, Oxford, United Kingdom of Great Britain and Northern Ireland; e-mail: wael.albayaydh@cs.ox.ac.uk; Helena Webb, School of Computer Science, University of Nottingham, Nottingham, United Kingdom of Great Britain and Northern Ireland; e-mail: helena.webb@nottingham.ac.uk; Omer Gunes, Department of Computer Science, University of Oxford, Oxford, United Kingdom of Great Britain and Northern Ireland; e-mail: omer.gunes@cs.ox.ac.uk; Marina Jirotka, Department of Computer Science, University of Oxford, Oxford, United Kingdom of Great Britain and Northern Ireland; e-mail: marina.jirotka@cs.ox.ac.uk.

ACM 2832-0565/2025/08-ART8
https://doi.org/10.1145/3737883

ACM Reference Format:
Towera Moyo, Wael Albayaydh, Helena Webb, Omer Gunes, and Marina Jirotka. 2025. A Responsible Research and Innovation Approach to Assessing the Implications of Anti-Trafficking Technologies. *ACM J. Responsib. Comput.* 2, 2, Article 8 (August 2025), 30 pages. https://doi.org/10.1145/3737883

1 Introduction

Responsible Research and Innovation (**RRI**) [27] is an approach that helps identify potential risks to individuals and society associated with new technical developments, as well as establish measures to mitigate their adverse effects. This approach emphasises the importance of engaging a wide range of stakeholders throughout the research and innovation process to thoroughly evaluate possible impacts and formulate strategies to maximise positive outcomes while minimising harm [27, 42, 59, 78].

The primary goal of RRI is not to predict the precise outcomes of an innovation but rather to anticipate and assess its various consequences, both intended and unintended. According to Franco-Santos and Otley's work, unintended consequences are defined as "the outcome of a conscious action other than that foreseen and intended" [32]. RRI process occurs in the early stages of the innovation cycle, ensuring that mitigation measures can be proactively incorporated into the development stage, whenever possible. As new technologies are constantly emerging to address various challenges, it is increasingly crucial to adopt such an approach to ensure that the technologies produce responsible outcomes. When rigorously followed, this approach helps ensure a coherent and responsible pathway for technological advancements.

There are no specific guidelines for the implementation of RRI in research or innovation projects [42]. However, Stilgoe et al. [74] developed a comprehensive framework to support and implement RRI in the research process. This framework was adapted and reworded by the **United Kingdom (UK)** Engineering and Physical Science Research Council, in collaboration with Stilgoe et al., and is known as the AREA framework [78]. The AREA framework is widely used in the UK [90] and consists of four key components: Anticipation, Reflection, Engagement, and Act [58, 62]. This framework encourages innovators to anticipate the intended and unintended consequences of technologies, reflect on the motivation of research or innovation, engage with various stakeholders, work together to anticipate and propose solutions to negative consequences, and take action by incorporating best practices.

Our study draws on the AREA framework to evaluate digital tools that are aimed at combating human trafficking. Although anti-trafficking tools that detect human trafficking offer numerous advantages, it is also crucial to examine and address the potential risks and potential negative consequences of these tools [19, 44, 45, 57]. By identifying the potential risks associated with anti-trafficking tools, it is possible to develop strategies to mitigate these risks at the beginning of the design process [45, 57]. This research was carried out through expert interviews and incorporated anticipation and engagement practices of the AREA framework to facilitate the examination of both the intended and unintended consequences of anti-trafficking tools. These consequences provided valuable insights, which in turn guided the development of design recommendations that can guide the process of developing anti-trafficking software.

The article begins with an exploration of the existing literature on technology-facilitated trafficking and the impact of digital tools employed to detect trafficking activities, which is elaborated in Section 2. In Section 3, we discuss the methodology and provide a background on RRI and its integration into the study. Section 4 presents the findings of our interviews with stakeholders. Finally, the article concludes with a discussion of the application of RRI in the field of **Information and Communication Technologies** (**ICT**). Our research focuses on digital tools designed

to identify victims and traffickers in cases of sex trafficking, as these tools are commonly used in anti-trafficking initiatives, and sex trafficking is the most prevalent form of trafficking [57, 82].

Our study makes several contributions to the academic community:

(1) We illustrate how the principles of RRI can guide the process of data collection and analysis, with a focus on examining unintended consequences.
(2) We demonstrate the application of RRI within the ICT domain, specifically in the evaluation of existing software systems.
(3) Our research provides novel empirical insights into the understudied area of anti-trafficking tool risks.
(4) We offer a set of design recommendations that various stakeholders can use when developing anti-trafficking tools.

2 Background and Related Work

In this section, we provide background information on technology-facilitated trafficking and anti-trafficking tools, which serve as crucial context for our research study and findings.

2.1 Technology-Facilitated Trafficking

The widespread use of Internet technologies has transformed the way human traffickers exploit victims. The terms "online trafficking" and "technology-facilitated trafficking" have emerged to describe the use of social media platforms, websites, and other online platforms by traffickers to recruit and exploit individuals [14, 16, 44, 45, 57, 82, 83].

In this article, we discuss three primary types of online platforms used by traffickers to carry out their operations. Each platform is described as follows:

— **Online classified websites**: These platforms serve as venues for users to publish **advertisements (ads)** for a variety of categories, including goods, services, and personal connections. On some platforms, sections specifically designated for adult services have emerged as hotspots for sex trafficking [44, 46].
— **Adult websites**: These websites advertise sexual activities and connect clients with people who offer sexual services. Although these platforms facilitate consensual transactions between consenting adults, they also provide a means for traffickers to communicate with victims and clients [46, 50].
— **Social Networking Sites**: These sites, also known as social media, allow users to establish connections, share information, and develop networks [44, 68]. However, these platforms also allow cybercriminals to exploit and target victims using personal information posted by millions of users [4].

These online platforms have proven to be highly advantageous for traffickers, as they provide access to individuals from a wide range of geographical locations, allowing them to recruit and exploit more victims than traditional means. The most widely used anti-trafficking tools are for the purpose of identifying victims or traffickers (Victim/Trafficker Identification tools), followed by tools for awareness raising, education, and collaboration (Awareness Raising, Education and Collaboration tools) [57]. This study specifically concentrates on Victim/Trafficker Identification tools, as mentioned previously. Throughout this article, Victim/Trafficker Identification is typically referred to as "tools" or "anti-trafficking tools". The following section discusses how anti-trafficking tools attempt to address the challenges of technology-facilitated trafficking.

2.2 Anti-trafficking Tools

In response to the increasing use of technology by traffickers, a variety of stakeholders, including law enforcement agencies, governments and non-governmental organisations, have begun to use technology to combat trafficking. Traffickers and victims often leave critical information in the online content they create, including keywords that can potentially aid in their identification. The tools for victim/trafficker identification employ **open-source intelligence** (**OSINT**) methods to analyse the data. OSINT encompasses a wide range of publicly available sources, such as social media platforms and escort ads [14].

Anti-trafficking tools have been used to analyse online content using various methods. Some of these tools use machine learning to analyse text in escort ads or social media content to identify sex trafficking indicators [23, 41, 86]. These indicators can take the form of keywords or phrases commonly employed by traffickers and buyers [82]. However, traffickers often employ deceptive tactics, such as altering keywords or employing obfuscation techniques, to avoid detection by machine learning models [77]. The use of such evasive measures by traffickers can hinder or prevent the accurate identification of trafficking ads.

Social Network Analysis (**SNA**) is also employed by anti-trafficking tools to pinpoint traffickers and their organized networks [3]. Relationships and connections between individuals on online platforms can be scrutinised using SNA to reveal the structure and dynamics of trafficking networks. It is crucial, as Campana [14] points out, to integrate indicators with SNA and expert knowledge, as relying solely on indicators may have certain limitations.

Certain tools employ image analysis techniques to identify potential victims. These tools analyse visual cues and patterns within images associated with online ads to detect signs of exploitation. Some of these tools identify potential victims by comparing images from escort ads or social media with photographs of missing persons [61]. In other instances, phone numbers and locations can be extracted from images and compared with known numbers connected to traffickers [57, 69]. In addition, some of these tools use visual cues and patterns to identify children in escort ads. The participation of children in commercial sex, whether willingly or not, is considered sex trafficking [80, 81]. Automation of the image analysis process has proven to be advantageous in terms of time savings because manual examination can be prone to errors and biases [61].

Various stakeholders use Victim/Trafficker Identification tools to meet their individual roles and objectives. These stakeholders include anti-trafficking organisations, law enforcement agencies, and individuals who have been affected by trafficking. Law enforcement agencies are the primary users of Victim/Trafficker Identification tools, and use them to support their investigations. Anti-trafficking organisations also employ these tools to rescue victims and report cases [57].

Similarly, developing anti-trafficking tools necessitates the coordination of multiple stakeholders such as the private sector, non-governmental organisations, government entities, and academic institutions. However, there are concerns about the motivations for the development of these tools, particularly when they are initiated by individuals or private entities [54]. Although the motivations behind the development of certain private sector tools may not always be apparent, these tools have had a profound impact on the operations of law enforcement agencies in both the United States (US) and the UK [61]. The following section provides details of the impact of these tools.

2.3 The Impact of Anti-trafficking Systems (Tools)

Anti-trafficking tools automate the data crawling and analysis process, significantly reducing the time required for analysis and minimising the likelihood of errors. These tools are crucial for

summarising potential trafficking cases for further investigation. Examples of such tools include Spotlight, PhotoDNA, and TrafficJam. For example, Spotlight has been instrumental in identifying more than 17,000 children in recent years. PhotoDNA has helped Google remove 270,000 users involved in child sexual abuse materials [30]. Furthermore, TrafficJam is estimated to have saved more than 70,000 hours of investigation time in 2020 alone [48].

Although anti-trafficking tools provide significant value, multiple studies have raised concerns regarding the potential bias and privacy issues associated with their use [18, 44, 45, 61]. The integration of sex worker data with trafficking victim data during data collection, processing, and storage raises significant privacy concerns [18]. Data breaches involving sex worker data can reveal personal information for sex workers and their clients, potentially leading to discrimination, stigmatisation, and threats to their physical and emotional well-being [19, 36]. There is a particular concern when entities such as law enforcement have access to such data, which substantially increases the vulnerability of sex workers [36, 75].

Anti-trafficking software can also expose victims to potential risks. Data mishandling can have severe consequences, such as the public disclosure of victims' personal information, leading to embarrassment and emotional distress, which can adversely affect their recovery process [19, 35, 36]. Several anti-trafficking tools store information on vulnerable individuals, such as children, thereby increasing the risk of harm and amplifying its impact.

Several studies have highlighted the importance of incorporating privacy considerations into anti-trafficking tools and investigations [18, 45, 57, 61]. The debate surrounding privacy versus security arises because strategies, such as encryption, may favour traffickers by enabling them to evade detection. Privacy laws further complicate the identification of victims and traffickers using such tools [14]. Addressing these concerns requires ongoing research, policy discussions, and collaboration to develop practical tools that respect privacy while effectively combating trafficking. More research is required to explore and develop methodologies that balance the identification of trafficking activities, while protecting the privacy of individuals involved in the sex industry.

A study by Musto [54] also highlighted reservations regarding the use of anti-trafficking tools. They stated that anti-trafficking tools are not developed based on accurate assumptions, as they are modelled after erroneous conceptualisations of sex workers, migration, and precarious labour. There was also scepticism regarding the potential benefits of these tools for victims, as it is believed that they are oriented towards favouring the private sector. Furthermore, they stated that these tools may cause more harm by diverting resources from other efficient solutions. Such issues may contribute to the scepticism faced by various organisations and individuals. They advised anti-trafficking organisations to thoroughly evaluate technology solutions before implementing them in practical settings to help ensure that they primarily benefit the victims. In addition, there should be no ulterior motives for the development and deployment of these tools [44, 45].

Most of the issues and concerns associated with anti-trafficking tools can be adequately addressed before or during the design of these tools. Deeb-Swihart et al. [19] calls for more ethical development of anti-trafficking and states:

> "Responsible design calls for researchers to directly engage with how AI impacts society and encode values that align with social norms. This highlights the need for researchers to center their design around considerations for potential harms and benefits of their work. This ties into the next principle of "considerations for long-term effects" as researcher should consider the benefits and harms beyond the immediate. Long-term effects of research are difficult to evaluate as academic articles represent snap-shots."

Responsible design should be implemented not only by researchers, but also by all stakeholders involved in the development of anti-trafficking tools.Stakeholders must identify methodologies to assess the long-term impacts of anti-trafficking tools and ensure that their designs are informed by these insights to mitigate potential harm.

Taylor and Latonero [75] also emphasise the importance of raising awareness of the risks and harms associated with technology and developing policies to safeguard human rights in the anti-trafficking domain. These policies should be established at the beginning of a project to ensure that harm mitigation measures are in place.

Although several studies have highlighted the significance of understanding risks and establishing guidelines for tool development, only a limited number have thoroughly examined their associated concerns [19, 44, 45, 57]. Most research has focused on privacy risks, particularly those affecting victims and sex workers; however, there are other potential negative impacts that need investigation. Therefore, it is essential to identify these risks and implement preventive measures [45, 57]. Many studies have overlooked the ethical implications or failed to propose methods to mitigate them. More comprehensive measures are needed to address data protection and ethics in the development of anti-trafficking tools [57], as inadequate safeguards could lead to adverse consequences for society, including harm to victims.

Based on the identified research gap concerning the lack of insight into long-term impacts of anti-trafficking tools and the absence of appropriate measures, we designed a study to address this question: "What are the potential consequences of anti-trafficking software and how can awareness of these consequences inform design guidelines in the US and UK".

2.4 Investigating Risks of Technology

The potential risks associated with anti-trafficking software are part of a larger concern about the risks associated with innovation. Poel and Royakkers [63] defined risks as the potential harm resulting from an undesirable outcome, taking into account both the likelihood of its occurrence and the severity of its impact. These risks can range from business-related concerns, such as safety issues, to ethical risks, such as reputational damage [8].

One of the primary risks posed by these systems is a breach of privacy, as highlighted in the previous section. Privacy is defined as an individual's ability to control their personal information and prevent its misuse [8]. The majority of innovations inherently carry privacy risks with the potential to adversely affect various societal groups. For example, concerns are frequently raised about the lack of transparency in information systems, which impedes the understanding of how data are collected, processed, and stored [24]. Common issues include the surveillance of personal data, misuse of collected information, and unauthorised access to protected personal data [8, 28, 39, 47].

Breaches of privacy create vulnerabilities, with significant societal consequences. For example, Zuboff [39] states how corporations such as Google exploit user behaviour data without informed consent, resulting in the commodification of personal information. This practice undermines individual autonomy and increases the risk of harm from data misuse [71]. Privacy risks have evolved into ethical issues as they threaten fundamental human dignity [31], highlighting the urgent need for robust regulations to protect individuals and society from the far-reaching impacts of surveillance capitalism.

Creators of innovations are responsible for preventing potential harm and undesirable outcomes [63]. Innovators have an ethical obligation to understand not only the technical risks associated with their creations but also the broader social, environmental, and human implications of these risks. They are accountable for mitigating these risks before their innovations enter the public domain. In the absence of appropriate guidelines, risks and potential harm may emerge during

the development of innovations [74]. This highlights the need for principles that help engineers identify these risks and ensure that technologies are developed in accordance with guidelines that prioritise societal, environmental, and human considerations.

Technology ethics, introduced to address these issues, aims to reduce harm, manage risks, and guide regulations and policies [37]. It has provided several benefits, including raising awareness of ethical concerns among stakeholders, such as engineers and policymakers, establishing ethical principles to guide technological development, and enhancing public trust in innovations. For example, the ethical guidelines developed by UNESCO [79] and the EU [26] aim to minimise the harm caused by AI systems while promoting transparency, fairness, and accountability. In addition, measures such as privacy frameworks have been created to help understand and preserve privacy [71].

Despite these achievements, technology ethics has been criticised for its narrow focus on individual engineers or design decisions, which often do not drive substantial change. Furthermore, it has been faulted to prioritise high-level principles over practical actionable rules [37].

Various stakeholders have implemented frameworks to enforce technology ethics and address the potential harm caused by innovation. The prominent frameworks include RRI, **Value Sensitive Design** (**VSD**), and **Technology Assessment** (**TA**), each designed to address these concerns [70].

VSD constitutes a theoretical and methodological approach that integrates human values in the design of innovations [65, 87]. It employs three types of investigation: conceptual, empirical, and technical to ensure that these values are effectively embedded [34]. TA focuses on evaluating the potential impact of technological innovations, ensuring that stakeholders such as policymakers and the public are adequately informed about their consequences [29]. As previously mentioned, RRI seeks to identify the potential consequences of technological innovations and propose measures to mitigate any negative consequences [27].

Although these frameworks share common objectives, they differ in scope and focus. RRI incorporates ethical and societal considerations throughout the entire innovation lifecycle. By contrast, VSD concentrates on these aspects during the design and development stages, offering a more specific focus [70]. TA focuses primarily on post-design phases and decision-making processes.

Anti-trafficking software operates within complex societal and ethical contexts, affecting various individuals across society. The primary objective is to ensure that stakeholders consider various implications throughout the innovation process. In light of the research question, which seeks to identify and understand potential risks and use this knowledge to guide the development of tools, incorporating established frameworks, such as RRI and TA, can be particularly valuable in addressing the research question.

RRI is particularly well suited for this work due to its emphasis on developing innovations that serve the public interest while aligning with societal needs and ethical standards [27, 42, 59, 78]. It also prioritises incorporating the perspectives of various stakeholders, including affected communities, and adopts a reflective approach that enables the innovation process to adapt to emerging challenges. Furthermore, its focus on the entire innovation lifecycle ensures that ethical concerns are considered at all stages. For these reasons, this article explores the application of RRI to address the research question. Additional details on RRI are provided in the following section.

2.5 RRI

Several RRI frameworks have been developed to operationalise this concept, as summarised in Table 1. The Directorate-General for Research and Innovation (DG RTD) of the European Commission and the RRI Tools project each established conceptual frameworks that encourage researchers and innovators to ensure that research and innovation are ethically sound, inclusive,

Table 1. Overview of RRI Frameworks

Creator	Year	Pillars
DG RTD	2013	Public Engagement, Gender Equality, Open Access, Ethics and Governance, Science Education, and Policy Integration
RRI Tools Project	2013	Public Engagement, Gender Equality, Open Access, Ethics, Science Education, and Governance
Stilgoe et al.	2013	Anticipate, Inclusion, Reflexivity, and Responsiveness
UK Research and Innovation	2013	Anticipate, Reflect, Engage, and Act
Jirotka et al.	2017	Anticipate, Reflect, Engage, Act, and Plus Component

and address long-term challenges [26, 67]. These frameworks improve the probability of public acceptance by highlighting public engagement, gender equality, open access, and risk mitigation.

The frameworks developed by Stilgoe et al. [74] and UK Research and Innovation [78] focus on four key areas: anticipating potential immediate and long-term harm, reflecting on current research processes, engaging a diverse range of stakeholders, and developing actionable steps to guide the research and innovation process. Despite these similarities, Stilgoe's framework presents more of an academic model of RRI that emphasises anticipating societal needs, fostering public dialogue, and providing comprehensive principles and guidelines for ethical practice.

The AREA framework diverges from Stilgoe's framework, as well as those developed by the DG RTD and the RRI Tools project in its pragmatic approach and emphasis on real-world applications [26, 67, 74, 78]. It was specifically designed to anticipate potential risks and mitigate adverse outcomes. The framework provides actionable steps for integrating responsible practices while prioritising simplicity, rendering it particularly suitable for developing responsible emerging technologies. Although other frameworks provide essential conceptual guidance for research and innovation, the practical approach of AREA is especially valuable when evaluating software with the objective of minimising harm in both the short and long term, making it suitable for this current study.

However, despite the benefits of the AREA framework, the practical application of RRI in the ICT field can present challenges for researchers and innovators, as RRI including AREA framework do not provide explicit instructions. Furthermore, the unpredictable nature of risks in the ICT field and the complexity of projects involving multiple stakeholders make the operationalisation of RRI particularly challenging [42, 85].

To address these challenges, Jirotka et al. developed the AREA Plus framework, which builds on the AREA framework specifically for ICT projects [42]. This enhanced framework incorporated an additional component that offered tools and resources to support each stage of the four AREA components. Researchers who encounter difficulties in implementing the AREA framework in ICT projects may find the AREA Plus framework particularly advantageous. The study employs the AREA framework because of its simplicity, practicality, and emphasis on anticipating consequences. The goal was to explore other frameworks, such as AREA Plus, if the study was unable to

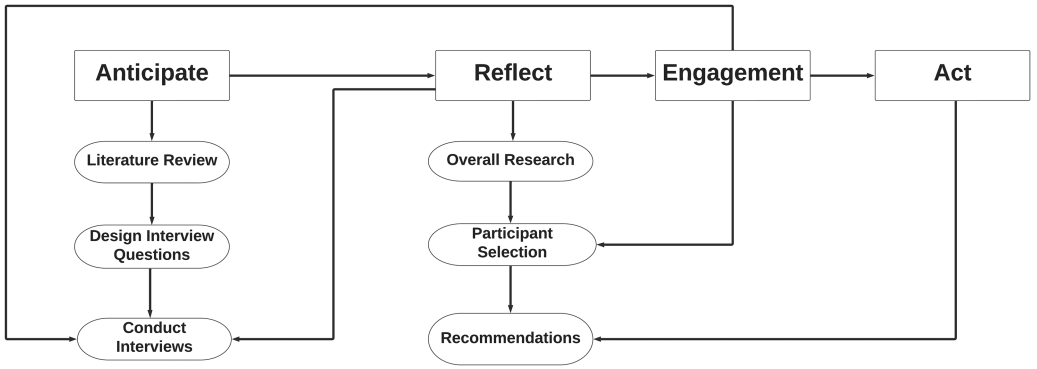

Fig. 1. Methodology: AREA framework.

anticipate the consequences; however, this was not necessary. The subsequent section discusses how this study used the AREA framework to address its objectives.

3 Methodology

Each component of the AREA framework contributes to a comprehensive understanding of the ethical, social, and environmental implications of research and innovation projects. The framework is deliberately non-prescriptive, and its components can be incorporated into research in multiple ways; for instance, sequentially, in overlap, singly, or iteratively. Figure 1 presents a summary of the use of the AREA framework in our research. A more comprehensive description can be found in the subsequent section.

3.1 Anticipate: Analysing the Consequences of the Tools

Anticipation is an essential aspect of RRI that involves considering the possible consequences and impact of a research project or innovation [56, 74]. This involves a thorough analysis of how a project can affect society, allowing the identification of challenges and risks [62]. Furthermore, anticipation encourages the formulation of potential solutions to address such challenges and risks [56]. The anticipation phase should involve relevant stakeholders to ensure that their perspectives and concerns are considered. Researchers typically interact with stakeholders through various methods, such as focus groups and interviews, to identify potential risks and implications associated with their projects [62].

This research project focuses on the anticipation aspect of the AREA framework, the primary objective of our study being to understand the implications and risks of Victim/Trafficker Identification, as previously stated. To achieve this goal, we conducted semi-structured interviews and a review of the literature. Semi-structured interviews were used because of their adaptable nature and the interviewer's ability to request additional information from the interviewees based on their responses [6]. This allowed us to obtain rich data on areas of interest. Interview questions were developed based on the anticipation stage. They were designed to get feedback from participants about the potential consequences, both positive and negative, and the risks associated with the tools and recommendations to mitigate the negative implications of the tools. We conducted a literature review to acquire background information on the topic and to supplement the information provided by the interview participants. The findings presented in Section 4 provide an overview of the anticipated consequences and risks associated with anti-trafficking tools.

3.1.1 Review of Literature. We conducted a search for literature related to anti-trafficking tools on various platforms, primarily Google Scholar and a university library catalogue. The latter includes non-print legal deposit books, scores, journal articles, and issues, as well as the UK Web Archive. Our search was not limited to any specific library, and we downloaded articles from any library containing relevant information on anti-trafficking tools. We employed several keywords in our search, including "sex trafficking", "human trafficking", "technology-facilitated trafficking", "online trafficking", "human trafficking Internet", "technology trafficking", "escort ads online human trafficking", "detecting human trafficking", "detecting sex trafficking", "online sex trafficking", and "child sex trafficking".

Snowball sampling was used to identify additional relevant literature. This approach involved identifying articles in the reference sections of the articles that we read. The reviewed literature allowed us to gain a comprehensive understanding of sex trafficking, technology-facilitated trafficking, anti-trafficking tools, and their consequences. This information complemented the findings of the interview.

3.2 Reflection: Critical Examination of Motivations and Benefits

The Reflection component urges researchers to extend their focus beyond traditional considerations and critically evaluate the motivations and benefits of their projects [56, 74]. This involves assessing whether research is necessary and whether its advantages outweigh potential risks. Researchers use reflection as an inward-focused approach to improve their informed decision-making, identify key stakeholders involved in the project, and foster the contemplation of broader questions related to the project at hand [62].

Reflection was used at various stages of the research. Initially, it was employed to assess the relevance and potential value of the study. Furthermore, we used reflection in this phase to guide the selection of key stakeholders to interview, paying attention to how their diverse perspectives could benefit the research results and add different voices. Victims of human trafficking (survivors) were not directly included in the research as participants because of concerns about the potential for (re) traumatising them by recounting their experiences of trafficking. Additionally, there was a lack of support for survivors during the research process in the event that welfare support was required. Instead, we identified ways to represent the voices of victims indirectly through participants with practical experience working with victims.

Reflection was also instrumental in formulating interview questions. The researchers were able to reflect on which questions would be suitable for different groups of participants and would yield significant information for the research. In addition, reflection played a crucial role in guiding the formulation of recommendations and design guidelines based on our findings. While devising the guidelines, the potential advantages and disadvantages of the suggested suggestions arising from the study were carefully considered.

3.3 Engage: Involving Diverse Stakeholders

The Engagement component involves the participation of relevant stakeholders such as experts and the general public [62, 74]. This approach benefits from a broader range of insights and ideas when considering various perspectives [78]. Collaborating with a diverse group of individuals, including academics, industry experts and social scientists, is an effective way to engage with stakeholders and gain valuable insights, feedback, and guidance that can influence the direction of a project and the decision-making process [62].

In this study, engagement was factored into the participant selection process to anticipate possible consequences. We intentionally sought input from a diverse range of stakeholders to ensure that the research considered multiple perspectives. The participants were law enforcement officers,

developers, and anti-trafficking experts. This group was chosen because law enforcement agencies and non-governmental organisations are the primary stakeholders that use Victim/Trafficker Identification tools in addition to developers who contribute to the development of these software tools [57]. By including a wide variety of voices, we ensured that all perspectives were considered while being sensitive to stakeholders' needs.

3.3.1 Recruitment and Interviews. Following the approval of our institution's Departmental Research Ethics Committee CS_C1A_021_017_001 [Approval: 17-09-2021] , we approached 28 candidates for participation and conducted 13 interviews. Three participants from our professional network were recruited via convenience sampling, while two additional participants were recruited through snowball sampling. A participant was recruited by e-mail. The remaining 19 individuals were recruited through a comprehensive search of experts from academic articles, previous well-known works, and LinkedIn. Of these, eight participants were approached via e-mail and the remaining 11 participants were contacted through LinkedIn.

Developers and authors of articles discussing the development of anti-trafficking tools were primarily identified through literature reviews. If institutional e-mails were not available in the articles, LinkedIn was used as an alternative. Law enforcement officers were recruited through existing networks and snowball sampling, and an officer was identified through a LinkedIn search. Keywords such as "law enforcement human trafficking" and "police human trafficking" were used to identify potential participants in this group. Anti-trafficking experts were identified through a combination of existing networks, snowball sampling, and reading of published works.

The recruitment process began with the distribution of an information sheet that describes the general description of the study, including the purpose of the invitation, potential benefits and risks, and data handling procedures. The information sheet assured the participants that their personal identifiable information would not be shared with any third party beyond the research team. Additionally, participants were notified of their right to withdraw from the study within two weeks of participation. Following the participants' agreement to the interview, a consent form was provided. For participants found on LinkedIn, the correspondence continued via e-mail.

The selection criteria for participants were as follows:

— To consent to participating in the interviews.
— To give consent to being audio-recorded.
— For developers, they needed to have experience with anti-trafficking systems' development
— For law enforcement, they needed to have experience with human trafficking investigations and anti-trafficking systems. We accepted participants who are retired but have previously worked on human trafficking investigations before retirement.
— For anti-trafficking experts, they needed to have experience in working in anti-human trafficking space.

The recruitment and interview of participants for this study proved to be a challenging task, particularly when securing the cooperation of law enforcement personnel. Consequently, we also extended our invitation to retired law enforcement officers who had prior experience working with anti-trafficking tools and who remained active in the anti-trafficking domain through alternative channels. The recruitment and interview processes spanned a period of six months.

Table 2 presents the characteristics of all the participants, including their job status, technological background, and experience with anti-trafficking efforts. We refer to participants as "P" followed by their participant number (e.g., P01 refers to Participant 1). The interviews were conducted using Microsoft Teams and a standalone recording device was used to capture the audio. The interviews lasted between 25 and 60 minutes, with an average duration of 45 minutes.

Table 2. Demographic Information of Participants

Characteristics	Number	%
Total Number of Participants	13	
Gender of Participants		
Female	6	46.2%
Male	7	53.8%
Job Type		
Anti-Trafficking Expert in Organisation	2	15.4%
Criminologist	1	7.7%
Detective Constable	1	7.7%
Retired Police Officers	2	15.4%
Computer Scientists	4	30.8%
Anti-Trafficking Expert/Founder	3	23.1%
Work Experience		
< 5 Years	3	23.1%
5–10 Years	4	30.8%
> 10 Years	6	46.2%

3.4 Act: Integrating Stakeholder Perspectives

In the AREA framework, the outcomes of the anticipation, reflection, and engagement components are used to identify opportunities to shape the trajectory of research and innovation in a positive way, which is the Act component [62]. In this stage, the insights and recommendations derived from stakeholder participation can provide guidance for ongoing project design and implementation [62].

In this study, the primary researcher and interview participants suggested a set of recommendations for anti-trafficking stakeholders when developing and using anti-trafficking software. These recommendations are intended to serve as a practical compass for future work on the design and development of Victim/Trafficker Identification Tools. The recommendations are based on the anticipated negative consequences of such tools and aim to mitigate them.

Following an RRI approach, we took care to be inclusive and respectful in our research activities, while also pursuing outcomes that can make positive instrumental contributions to support good design practices for anti-trafficking tools.

3.5 Analysis

The primary researcher transcribed and coded the interviews using an inductive approach that involved identifying codes from the data. The initial analysis conducted after each individual interview indicated that data saturation was observed after the 12th interview, with no significant new codes emerging (from interviews 13 onwards). For data analysis, researchers have adopted Thematic Analysis [10]. Thematic analysis is a well-established approach to explore areas with limited prior research, making it suitable for this study.

One hundred and twelve codes were identified and organised into themes using axial and selective coding. Following an initial phase of inductive and exploratory analysis, we drew on the

Table 3. Themes and Sub-Themes

Theme	Sub-Themes
Positive consequences of anti-trafficking tools	1. Positive societal changes
	2. Aids in investigations
	3. Prevents people from working with explicit content
Unintended negative consequences of anti-trafficking tools	1. Victimisation of vulnerable people
	2. Negative consequences for victims
	3. Social consequences of misclassifying traffickers
	4. Privacy implications of tools
	5. Adversarial efforts of traffickers
	6. False sense of security
Recommendation to mitigate negative consequences	1. Design considerations
	2. User considerations

AREA Framework, focussing on codes and analytic insights related to the various consequences of anti-trafficking tools and mitigation measures for negative consequences.

Braun et al. [11] warned against relying solely on data saturation in thematic analysis and advised researchers to consider contextual and practical information, such as research questions and data richness. This study carefully designed the research questions to obtain detailed data. Furthermore, the diversity among the participants contributed significantly to the richness of the data.

3.5.1 Comparative Approach: US and UK Contexts. This study analysed Victim/Trafficker Identification tools in the US and the UK. Most anti-trafficking tools have been developed in the Global North [57], and the majority of these tools and research work have originated in the US. Therefore, the inclusion of the UK in this study provided a comparative approach to examine the implementation and impact of these tools in both jurisdictions. This approach allowed us to discover potential similarities and differences in the results by employing a comparative approach, providing a more robust understanding of the effectiveness and implications of these tools. The following section presents our research findings.

4 Research Findings

The analysis of anti-trafficking tools has revealed a variety of anticipated consequences, each presenting distinct levels of harm or benefit. In addition to these consequences, the participants provided recommendations to address potential negative concerns. Table 3 presents an overview of the themes and sub-themes that emerged from our investigation. While recognising that the effects may differ between the US and the UK, each theme's discussion clarifies the extent of similarity or disparity where applicable.

4.1 Positive Consequences of Anti-trafficking Tools

This section discusses the positive impact of tools on users, recipients, and society as a whole. The results of this analysis revealed that most of the positive outcomes were anticipated and aligned with the intended objectives of the tools. Although some positive consequences were unforeseen, they were proven to be significant. Furthermore, the tools showed parallel positive effects in both the US and the UK.

4.1.1 Positive Societal Changes. According to the participants, these tools have provided a more comprehensive understanding of human trafficking, enabling the allocation of financial and non-financial resources to combat sex trafficking. These resources have significantly benefited stakeholders, such as law enforcement and policy makers. For example, these tools have proven invaluable in identifying trafficking hotspots, allowing stakeholders to strategically channel their prevention and rescue efforts and make informed decisions regarding resource distribution. As the use of technology in trafficking cases continues to increase, the identification of these hidden locations has become increasingly challenging. One participant, P02, shared their unit's experience using these tools to locate victims:

> "We need to know where to look for them because they're now in ordinary houses and regular streets'in pop-up brothels, Airbnb rentals, or other short-term lets. It's much more hidden. Using adult service websites and web scraping, we can locate people and work to safeguard them." (P02)

Furthermore, the use of these tools has brought attention to global challenges, particularly exposing the deficiencies inherent in existing legislation. This has led various stakeholders to initiate proactive measures. Participant P01 noted that the laws themselves exhibit imperfections. Although this was not the intended outcome of tool development, the awareness sparked by these tools has had positive repercussions in certain countries, including the UK.

The participants also highlighted the significance of the intelligence provided by these tools in the formation of modern slavery units in the UK. These units have greatly influenced investigative strategies, moving from traditional police procedures where victim needs were not always first to approaches that prioritise victim needs. This evolution has proven to be particularly beneficial for people who have experienced exploitation.

These tools have played a crucial role in the discovery of previously unknown information and in the provision of valuable information to stakeholders. This new knowledge has allowed stakeholders to pinpoint areas that need improvement, which is beneficial in the fight against sex trafficking. Furthermore, these tools have aided in the distribution of resources and have been instrumental in assisting law enforcement agencies in conducting investigations. The use of tools such as PhotoDNA provides an example of how such tools are shaping legislative work. PhotoDNA was recently discussed during 2022 state hearings on Digital Responsibility in the US [22], specifically in relation to the EARN IT Act. This Act grants online service providers, such as Facebook, the ability to remove content related to sexual exploitation of children without facing legal repercussions from individuals [17]. The subsequent section discusses how these tools assist in the investigation efforts of law enforcement.

4.1.2 Aids in Investigations. Most of the participants acknowledged that these tools have played a crucial role in the analysis of large amounts of data found on online platforms. This is particularly beneficial for law enforcement agencies, which often struggle with limited time and responsibilities. These tools simplify the information gathering process, which would otherwise require substantial time and resources. Tools of this nature have the potential to significantly reduce the time spent on investigation, some reports suggesting a reduction of up to 50% [73]. Additionally, these tools are known to be up to 60 times faster than manual search methods [48].

These tools also aim to prevent law enforcement from relying solely on intuition to conduct sting operations, instead promoting a more informed and evidence-based approach.

In addition, the tools not only help combat sex trafficking, but also provide valuable information for law enforcement agencies to identify and prevent other criminal activities. For example, traffickers frequently engage in illicit activities, such as drug trafficking, which may come to light

through the use of these tools. P12 noted this unexpected benefit as a favourable outcome resulting from the use of the aforementioned tools:

> "They are a valuable resource because they help build information and intelligence about broader criminal activity. For instance, if an organized network is involved in sex trafficking, these groups are often poly-criminal. It's not just sex trafficking–they may also be involved in drug trafficking, money laundering, and other criminal practices." (P12)

Participants noted that human traffickers are often convicted of other offences because the trial of human trafficking can be challenging to prosecute. Victims may not always be willing to testify for a variety of reasons, including the fear of facing legal consequences. As a result, accumulating information on other criminal activities is advantageous, as governments can prosecute traffickers on those charges based on their involvement in those crimes.

Furthermore, the benefits of these tools extend to preventing law enforcement from directly engaging with explicit content, which also includes other stakeholders, such as content moderators and anti-trafficking experts. Working with sex trafficking content can be emotionally and psychologically challenging as individuals are exposed to explicit and potentially triggering materials. The use of tools plays a crucial role in addressing this challenge by automating processes and reducing the amount of explicit content that stakeholders would otherwise have to analyse manually. This reduction in workload significantly improved the well-being of those working in this field, and P12 recognised the challenges associated with working in this environment:

> "You're spending hours a day looking at highly sexually graphic content. While it's not necessarily violent–it's standalone pictures, not videos–it is still very explicit. Many of the things we tag involve ethnic slurs related to body descriptions or language that is extremely sexually explicit, describing certain acts being performed." (P12)

As tools are increasingly used to investigate trafficking, they simultaneously cause a displacement effect and compel traffickers to adopt novel tactics. The following section goes deeper into this unintended consequence as well as other unforeseen negative implications.

4.2 Negative Consequences of Anti-trafficking Systems

This section highlights the potential negative consequences of using anti-trafficking tools, as identified and discussed by the participants. It examines the challenges that can arise from using these tools, which are often influenced by existing issues with the tools themselves, as well as broader social, cultural, and legal factors. The findings suggest that while developers may not always consider or foresee these unintended consequences, other stakeholders with their experience and expertise in the field may be able to anticipate such issues.

4.2.1 Consequences of Misclassifying Traffickers. Participants expressed considerable concerns about the substantial rate of tool misclassification, as it can have severe implications for people who are mistakenly identified as traffickers or victims. The interviews revealed several factors that contribute to this misclassification, such as the scarcity of available data and the inherent restrictions of red flag tools. The subsequent section delves deeper into the specifics of these factors, leading to misclassification.

Limited data. Participants expressed concern about the scarcity of data in the anti-trafficking domain. A significant issue that has emerged is the need for ground-truth data to support technological solutions. Unfortunately, the lack of ground-truth data can be attributed to the low rate of prosecution in trafficking cases, which in turn impedes the acquisition of the required data for ML applications. According to the **Organisation for Security and Cooperation in Europe** (**OSCE**),

there was a global decline of 42% in trafficking prosecutions between 2015 and 2018 [57]. The author states that the decline in human trafficking prosecution can be attributed to the remarkably low risk faced by these criminals with respect to being identified, prosecuted, and sentenced. Furthermore, the fact that traffickers operate with impunity contributes significantly to the overall decrease in their prosecution rates.

Participants highlighted the causes of the low conviction rate for human trafficking cases. Courts often convict human traffickers of crimes other than the offence of human trafficking, which leads to an insufficient prosecution rate. This is a significant contributor to the scarcity of data on human trafficking.

Participants also expressed concerns about the lack of data sharing among stakeholders, which exacerbates the problem of limited data availability. The unwillingness to share data hinders other stakeholders from obtaining the crucial information required to design effective counter-trafficking strategies.

In addition, the participants attributed the high rate of misclassification to the use of red flag tools or rule-based systems. These mechanisms rely on predetermined indicators or patterns to identify potential traffic instances.

Limitations of Red Flag Tools. According to Campana [14], there are limitations in using indicators to identify trafficking, which several participants confirm, as they acknowledge that red flag tools often result in numerous false positives. These false positives can inadvertently create an environment that is favourable for human traffickers to operate. One participant emphasised that tools should only serve as a means of providing insight rather than relying solely on identifying potential victims or perpetrators.

A participant displayed complete scepticism about the usefulness of red flag tools. Specifically, Participant 13 expressed doubts about the reliability of keywords or phrases used by traffickers as indicators.

> "I think there is zero probability that forced prostitution has the key as a keyword tell or some flag. Nobody advertises their illegality, nobody. And if it appears they are, it is likely that they are misleading." (P13)

During the discussions, the participants highlighted various negative social consequences that could result from the misclassification of traffickers. False labelling of individuals as traffickers can lead to stigmatisation. P06, a law enforcement officer, shared their experiences with the stigmatisation associated with such a misclassification:

> "Sometimes, they can get it wrong. As we discussed, the social consequences of such mistakes can be significant. The stigma for an innocent individual wrongly identified as an offender can be incredibly heavy. So, yes, the consequences of getting it wrong are quite severe." (P06)

Stigma can have negative effects on individuals, including inducing feelings of shame and isolation [9]. Moreover, it can lead to more severe consequences, such as bullying, targeted violence, and reduced opportunities, which can subsequently affect one's financial circumstances [55].

Misclassification of traffickers can also lead to a breakdown of families if there are suspicions of human trafficking. Families may experience social stigmatisation and financial hardship while navigating legal proceedings, which can cause significant psychological distress. Furthermore, participants noted that the majority of the people affected are male, which may disadvantage women and children in heterosexual family structures. However, they also acknowledged that software is not always the primary cause of these issues. In certain instances, misinterpretation of the tool results contributes to misclassification.

The following section discusses how anti-trafficking software can lead to victimisation of vulnerable individuals.

4.2.2 Victimisation of Vulnerable People. The impact of these tools is influenced by the country in which they are deployed, as risks and benefits can vary depending on the country's existing laws, practices, and culture. Concerns about the potential victimisation of vulnerable individuals, particularly sex workers and people of colour, are often linked to countries such as the US, where sex work is illegal and where there may be a history of tumultuous relations between police and marginalised communities. In such contexts, these tools are likely to contribute to discriminatory profiling and exacerbate existing inequalities and injustices. The tools may lead to discrimination because of the data used for the system and the design.

Additionally, participants expressed concerns that the tools that typically use personal information provided by sex workers on escort or adult websites might lead police officers to approach them based on this intelligence. This, in turn, can result in arrest or other unfavourable consequences. A significant challenge identified by participants was the difficulty faced by tools and developers in distinguishing between voluntary sex work and sex trafficking. This lack of differentiation poses challenges to sex workers, as it increases the likelihood of incorrect classification, potentially labelling them as victims while engaging in consensual sex work. In countries where engaging in sex work is prohibited, such as the US, this confusion between voluntary sex work and human trafficking can also result in individuals being unfairly labelled as criminals, even if they are engaging in consensual sexual activities. P10 raised important ethical concerns about this misidentification:

> "The other issue is that, when you look at the red flags or the rules from other companies you've researched, they often confuse sex work with human trafficking. As a society, we need to decide whether that's acceptable. Is it okay to treat all sex workers as if they are human trafficking victims? I don't think all sex workers want to be treated that way. Many of the interventions I've seen, particularly the technologies, blur this distinction." (P10)

Despite the illegality of sex work in other countries, some participants expressed concerns that law enforcement officials might still approach sex workers based on the information provided by these tools.

Anti-trafficking tools can also be used by online platforms to monitor their platforms. However, there is concern that misclassification could result in individuals being banned from online platforms due to the **Fight Online Sex Trafficking Act** (**FOSTA**) and the **Stop Enabling Sex Traffickers Act** (**SESTA**) in the US. This Act holds platform providers responsible for information regarding prostitution, whether consensual or not, posted on their platforms [66]. Since the introduction of FOSTA-SESTA in 2018, several sex workers have been banned on online platforms [15]. Therefore, these tools may have unintended consequences for sex workers.

Furthermore, several participants voiced their concerns about the possibility of people of colour being subjected to victimisation, particularly in countries such as the US, where historical tensions between marginalised communities and law enforcement persist. They highlighted that individuals misidentified by the tools could suffer significant consequences, particularly in situations where racial profiling and discriminatory practices are prevalent. Facial recognition systems have demonstrated reduced accuracy when applied to individuals of colour, thereby potentially increasing the risks associated with anti-trafficking software that utilises facial recognition technology [19].

According to a UK law enforcement officer, it is not the practice of law enforcement officials to intentionally target individuals who are not victims. If this occurs, it is recommended that law enforcement refrain from such actions and implement protective measures for sex workers.

Additionally, there are concerns about the possible negative consequences that can affect real victims of sex trafficking, as detailed below.

4.2.3 Negative Consequences for Victims. Participants reported that traffickers often force people to engage in illicit activities such as visa fraud by duress, coercion, or deception. It is possible for a nation's legislative framework and operational procedures to fall short of protecting victims, leaving them vulnerable to prosecution. In the UK, the 2015 Modern Slavery Act provides protection for victims forced to commit such crimes [38]. However, according to a review of the literature and some participants, the US does not have a similar law in place to protect victims. P01 emphasized the necessity for legal measures that prioritize the protection of victims:

> "There is some protections that need to happen for the victim. Victims are committing crimes as part of their exploitation, and we need indemnification for those crimes." (P01)

The criminalisation of trafficking victims can impede their ability to improve their lives after being exploited, which may lead to their re-exploitation and perpetuate the cycle of victimisation. Furthermore, the challenges that victims face in reintegrating into society, such as finding employment, can discourage them from viewing rescue operations as viable options. Additionally, some victims may resist rescue efforts if they believe that their current situation is better than that of their past lives. As a result, victims may not always be open to rescue, and their resistance can increase the likelihood of their return to the trafficking ecosystem.

Certain tools have been designed to help victims. However, these types of tools can also pose risks to the individuals they are meant to assist, as stated by P05.

> "Another concern is what happens if someone develops a tool that displays something on a phone, like an ad for a hotline—for example, the National Human Trafficking Hotline. If a trafficker sees that on the phone of someone experiencing trafficking, it could lead to violence, punishment, or even death. Something intended to help, like an ad, could have dangerous consequences." (P05)

4.2.4 Adversarial Efforts of Traffickers. As the efficiency of law enforcement in processing cases improves, traffickers are attempting to maintain the continuity of their operations.

Traffickers have been known to develop innovative tactics to elude detection by tools, such as employing emojis in their ads instead of text to evade keyword searches or disguising crucial information on images to avoid detection by tools that do not use image detection technology. P09 shared their experiences with the adversarial efforts of traffickers, highlighting the need for ongoing adaptation and improvement of anti-trafficking tools to remain effective in the face of these challenges:

> "Another issue was that phone numbers were often crucial information, but sometimes the poster would include them in the image instead of the text or use something like a CAPTCHA. This made it much harder to access the number. You'd need computer vision algorithms to extract it, which complicates the process significantly." (P09)

The participants offered instances where traffickers had modified their strategies to avoid detection by anti-trafficking measures. One such instance involved the deliberate exclusion of important information, such as phone numbers, in ads to prevent the detection of victims' whereabouts by the tools. This adaptability of traffickers poses challenges to the effective implementation of technological solutions in the fight against trafficking.

The tactics employed by traffickers are constantly evolving, which can create a sense of fear and hesitation among anti-trafficking stakeholders when it comes to openly discussing or publishing

their works. There is a concern that sharing too much information about the technology solutions used in anti-trafficking efforts may inadvertently assist traffickers in understanding how the tools operate, leading to adjustments in their behaviour to evade detection.

Furthermore, the participants raised concerns about the relevance of the information used in the development of anti-trafficking tools. Traffickers often change their tactics and adapt to new platforms or communication methods, rendering specific keywords or phrases ineffective in identifying trafficking activities over time. As a result, these tools can become less effective in detecting and combating evolving forms of trafficking if they are not regularly maintained and updated. In addition to these issues, there are also concerns about the privacy associated with the use of these tools.

Additionally, traffickers frequently alter their online platforms to avoid detection, leading to the proliferation of criminal activity on various online platforms. This displacement effect presents a challenge when using anti-trafficking tools, because traffickers can disperse, making it difficult to identify and monitor [57]. As evidenced by Backpage [7, 40] and MyRedBook [54], this was the case. Backpage was a classified site where over 70% of sex trafficking in the US occurred; however, it was shut down by law enforcement, resulting in traffickers using other platforms, which makes it difficult to identify them [7, 40]. The closing of websites or sections of a website is a common action taken by law enforcement when they discover instances of significant human trafficking on online platforms.

The closure of digital platforms commonly used by sex workers can have negative consequences on their livelihoods, as many depend heavily on these platforms for their work. Unfortunately, the shutdown of popular platforms has led to an increase in street-based sex work and the control of sex workers by third-party individuals [25, 76].

4.2.5 Privacy Concerns. Several studies have expressed concerns about the use of data related to human trafficking, as this information is highly sensitive, and any mismanagement could result in severe consequences [18, 44, 45, 57, 61].

Participants expressed similar sentiments and concerns about the potential breach of privacy that may arise from the use of these tools. They were concerned about the collection of personal identification information through these tools, which, if connected, could provide insights beyond the intended scope of data collection. The participants pointed out that some people working in this field tend to unnecessarily collect personal information to identify victims and traffickers. P10 talked about the tendency of developers to grab data: *"The second is, is that I see technologies that violate privacy of people who are, there's no reasonable, there's no reasonable standard to grab their personal data" (P10)*

Furthermore, the participants expressed apprehension about the possible misapplication of the data obtained using these tools. They emphasised the level of intelligence that these tools can attain by linking diverse data points. They also expressed concerns about the intrusive nature of these tools, particularly in relation to sex workers. These tools collect and process personal information without consent, thus eliciting considerable privacy concerns.

As the use of online platforms, such as adult/escort websites, by sex workers has become more prevalent, these platforms are gathering and analysing an increasing amount of information and details. This raises ethical concerns regarding the extent to which these platforms respect privacy and provide consent to sex workers throughout the entire process.

In addition to voicing privacy concerns, participants suggested practical measures to alleviate the possible adverse effects of these tools. The following section describes these measures in more detail.

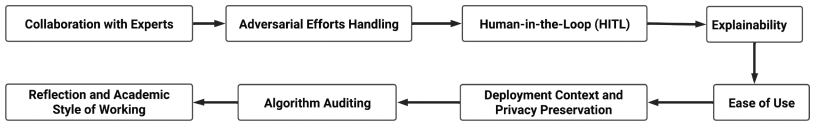

Fig. 2. Design considerations.

4.3 Recommendations to Mitigate the Negative Consequences

This theme presents recommendations from participants and previous research on strategies to proactively address identified issues and help alleviate negative consequences. These recommendations are geared towards the design and post-implementation phases of the tools.

4.3.1 Design Considerations. This section discusses the design considerations stakeholders should take into account; these considerations are summarised in Figure 2.

Collaboration with experts. Participants stressed the importance of involving people with a deep understanding of crime during the early stages of development. They recommended the formation of diverse teams to ensure a comprehensive examination of the tools. A specialised working group consisting of sex workers and survivors was proposed, with advocates standing in for them, as necessary. The views on the participation of survivors in the development of tools varied. However, most of the participants believed that survivors, particularly those with expertise in technology-facilitated trafficking, should be included in the design process.

Furthermore, as noted by Deeb-Swihart et al. [19], the participation of survivors can be crucial in vetting the ground-truth tools used to train AI models, thereby improving the accuracy of these tools. Focus groups can serve as an effective means of engaging survivors in developmental processes. Previous research [19] has advocated granting survivors and disadvantaged communities a sense of ownership over the project, either as project leaders or researchers, to achieve equitable participation.

To formulate a specialised working group for the development of victim/trafficker identification tools, it would be essential to also involve other key stakeholders, such as law enforcement agencies and non-governmental organisations. These connections can be formed through existing networks using professional platforms such as LinkedIn, reviewing relevant academic articles, and conducting searches on Google to identify organisations operating in the field [52].

Adversarial Efforts Handling. To effectively counter the evolving strategies of human traffickers, it is crucial to adopt a proactive and adaptable approach that integrates state-of-the-art technologies with human expertise. There are numerous methods to address adversarial efforts, and considerable time and resources should be invested in identifying suitable approaches for projects. Examples include adversarial training [5], transfer learning [91], and ensemble learning [21].

In addition to these techniques, we strongly advise developers to incorporate feedback loops into AI systems designed to combat human trafficking. Feedback loops allow developers to refine model elements after deployment by using the output of a model or user feedback [72]. As AI tools may make mistakes and mislabel information, feedback loops allow AI systems to recognise their errors and learn from new data to improve performance [13]. Retraining models with new data and adjusting hyperparameters are among the processes involved in feedback loops that help the model adapt to new changes.

Moreover, incorporating a human into the learning loop can be beneficial in recognising new tactics [88]. AI systems can better adapt to evolving criminal strategies by fine-tuning models with the help of human experts. There are several advantages to including humans in the process, as discussed below.

Human-in-the-Loop (HITL). The HITL method emphasises the active role of humans in the development process, from data collection and model training to evaluation and refinement. The study participants strongly recommended implementing an HITL approach for anti-trafficking software to ensure greater accuracy and mitigate the risk of false positives. This approach enables humans to directly influence the learning and decision-making processes of a machine learning system by providing input and decision making [51, 88, 89].

Human oversight is essential to ensure the accuracy and reliability of the information obtained from the AI tools. Including domain experts in final decision-making processes is crucial for mitigating the risk of false positives, which can have severe consequences. The AI model continuously learns from human input, and experts can participate in the process through various methods such as data annotation.

Incorporating an HITL approach improves interpretability and improves the understandability of model results, particularly in natural language processing tasks [88]. Explainability and interpretability of model outcomes in anti-trafficking tools are critical for preventing potential negative consequences. The following sections further discuss the importance of interpretability and explainability in the context of anti-trafficking tools.

Explainability. It is essential that the tool's decisions are easily comprehensible to users. Explainability plays a vital role in facilitating expert decision-making and assessing the reliability of algorithmic results. Incorporating explainability was strongly recommended by both participants and previous studies. The UK's Justice and Home Affairs Committee has expressed concern about AI systems that predict crime and the potential for unjust criminalisation based on unexplained intelligence [43]. The ability of users to understand algorithmic decision-making builds trust, which is crucial for both experts and society, particularly in situations where the insights derived from these tools are used as part of a legal defence to justify decisions [12, 33]. This trust is particularly relevant given the mistrust observed among some experts. However, achieving explainability can be challenging, particularly with complex models that are not always interpretable [12, 33]. Therefore, it is essential to devise methods that promote the interpretability of the models used in anti-trafficking software development. Furthermore, appropriate models for the problem should be selected, taking into account complexity, architecture, and data quality and quantity [64]. Easy comprehension of the model can also be enhanced by its ease of use, as explained in the following section.

Ease of Use. The significance of designing user-friendly tools has been highlighted by participants and in previous research [14]. Developers must create tools that cater to a diverse user base with varying levels of technical proficiency. Simplified tools can help reduce the risk of misinterpretation and cater to non-technology-savvy users. The development of user interfaces that enable seamless information flow can help address this issue. Collaboration with experts during development ensures that the needs, values, limitations, and abilities of users are understood. More research is needed to explore user interfaces tailored to the specific needs of different stakeholders.

We recommend that developers incorporate **user experience** (**UX**) frameworks or principles, such as UX honeycomb [84], to ensure positive UX. Ultimately, working closely with users is crucial to achieving usability of the system.

Deployment Context. Developers should investigate the deployment context during the design phase. The effectiveness of a tool varies based on the legal framework, cultural norms, and operational practices of the location where it is deployed. Customising the tool for the location is important because certain features may have unintended negative consequences in different settings. By considering contextual factors, including the legal and cultural nuances of each

country, developers can ensure the suitability and effectiveness of their tools in international deployment.

Privacy Preservation. Privacy protection is especially critical for marginalised groups, who face heightened risks; these are necessary measures needed to enhance their security and autonomy [1]. Privacy protection is critical for marginalised groups, who face heightened risks; measures are needed to enhance their security and autonomy. Given this, stakeholders must implement safeguards to ensure privacy during data handling, including collecting only the data necessary for the system's function. Systems should avoid gathering information unrelated to human trafficking. Where such data collection is unavoidable, steps must be taken to anonymise the data or remove personal identifiers [52]. Preventing the identification of victims should remain a primary priority for stakeholders [36].

Furthermore, robust measures should be implemented to safeguard data from breaches during the development and deployment stages. For instance, secure coding practices such as utilising separate computers for data scraping and processing can enhance data security and minimise the risk of unauthorised access. Developers should also conduct routine code reviews, assess vulnerabilities, adhere to best security practices, and ensure compliance with legal data processing requirements.

Furthermore, other machine learning privacy techniques and frameworks can be incorporated to enhance the security of anti-trafficking systems. For example, privacy-preserving models can be implemented, differential privacy frameworks can be adopted, and noise can be introduced into datasets to safeguard sensitive information [60]. Furthermore, machine learning techniques can facilitate dynamic adjustments to privacy settings after deployment, ensuring that the system remains secure and adaptable as data flows and usage patterns evolve over time [2].

Audit Trails for Supervision. Tools should be equipped with audit trails to prevent misuse and maintain accountability. Implementing a zero-trust mindset serves to protect against potential harm to users. Previous research [19] and participants have expressed concerns about the misuse of these tools by certain stakeholders when deployed, such as stalking ex-partners. An audit trail can be used to monitor their use and alert appropriate individuals to any suspicious activity. Some participants have implemented this feature in their software tools to help monitor the use of the tools.

Data Sufficiency. The quantity of data plays a crucial role in determining the efficiency of the tool, as highlighted by the participants in the previous section.

Following data collection, system developers should assess the volume and adequacy of data. If data is insufficient, they should consider exploring secondary sources or collaborating with stakeholders to acquire additional data. Alternatively, developers could generate synthetic data. However, this method requires thorough testing and validation. Furthermore, techniques such as data augmentation can be employed to enhance the quality and diversity of the datasets ([53]). Sufficient data can improve the accuracy, thereby addressing the issues of misclassification.

Algorithm Auditing. Auditing the algorithm involves evaluating the performance of the algorithms and, consequently, addressing potential social harms, such as biases [43]. Independent auditor evaluations provide valuable information in internal processes and documentation, thus increasing accountability and transparency [20, 49].

Metaxa and Hannock [49] proposed conducting algorithm audits for tools designed to create social impact. These audits address concerns raised by participants in the previous section, such as false positives. Audits contribute to system improvement by identifying the rate of false positives and analysing the scenarios that lead to such errors. Furthermore, auditing involves testing the

Fig. 3. User considerations.

system on diverse datasets to ensure accuracy and fairness, as well as identifying and mitigating biases. Taken together, these processes play a crucial role in reducing false positives and improving overall system reliability.

The authors recommend anticipating the consequences of performing algorithm audits and rectifying any associated problem [49]. The AREA framework [74] or the AREA Plus framework [42] can be used to assess the impact of these audits.

Reflection and academic style of working. When designing a tool, developers must consider several elements, including the nature of the data to be used and whether ample data is available. It is crucial for stakeholders to thoroughly assess these factors during the design process.

The participants highlighted several considerations for developers when designing tools to combat human trafficking.

- Determine how the tool will distinguish between voluntary sex work and sex trafficking.
- Ensure that the data are carefully selected and of good quality. The questions to answer include the following: Why are you collecting these data? Is the data suitable, diverse, and sufficient? Who will have access to these data? How can this data be used for other unintended purposes? How do you keep the data relevant? What will happen to the data over the next five years? How will you secure the data?
- Identify ways in which the tool will be kept relevant. The questions to answer include the following: How do you keep the tool relevant? How do you keep the tool searching for the right thing?
- The users of the tool should be thoroughly researched, and the design should be dependent on the users. The questions to answer include the following: Who will use the tool? What are the interests of the users? How can the tool be used by various stakeholders? How would you prevent the tool from being used for other unintended purposes?
- Consideration should be given to whether the tool will be implemented in different countries or cities and how it can be adapted to accommodate unique characteristics. The questions to answer include the following: How will the tool be implemented and integrated within the socio-economic, cultural, and infrastructural context of the country, city, or state of deployment? What strategies will be employed to address and mitigate the impact of false positives within the system's operation?

In addition, several participants suggested that the private sector should adopt practices commonly used in academia. Academics are renowned for their extensive documentation of their methodology, intentions, and justifications for the problem and their building on previous research. Establishing clarity in the plans from the outset is crucial for developers. This clarity should encompass the methodology they intend to employ, their underlying intentions, and the proposed solutions they plan to implement. Previous studies [54] and participants have indicated that the motivations for developing anti-trafficking tools are not always clear, and adopting an academic writing style helps ensure that the research problem has been thoroughly explored and the contributions are explicitly stated.

4.3.2 User Considerations. The participants offered recommendations for the responsible use of the tools by stakeholders; these recommendations are summarised in Figure 3.

Respect for Sex Workers' Rights. Stakeholders ought to focus on creating protective legislation for sex workers, collaborating with local sex worker charities, restricting private data access to authorised individuals, and adhering to existing laws, such as the UK's National Police Chief Council guidance on sex work.

Additionally, legislation must push tools to distinguish between voluntary sex work and sex trafficking. Such differentiation would be immensely beneficial for both anti-trafficking users and developers.

Enhancing Dataset Quality. Foster collaborative initiatives between anti-trafficking stakeholders by establishing various working groups for joint investigations and data sharing. Strong relationships can facilitate the development of effective solutions using shared, high-quality, and anonymised data.

Privacy. Access to tools should be regulated, granting access only to authorised individuals. It is essential to limit data access to only those with legitimate needs and promptly delete extraneous personally identifiable information to avoid the potential for misuse.

These steps can improve tool usage, while safeguarding privacy and promoting effective anti-trafficking efforts.

5 Discussion and Recommendations

This study was based on the AREA framework in RRI to evaluate the social impact of anti-trafficking tools and offer recommendations to mitigate the potential negative consequences associated with their use. The framework was instrumental in guiding our research design and analytic focus to assess the intended and unintended effects of these tools. The findings provide novel insights into the application of RRI principles in the evaluation of software systems, thereby demonstrating the practical utility of RRI in ICT studies. Furthermore, the findings also provide valuable empirical information in an understudied area and offer design recommendations for stakeholders to consider to mitigate the potential negative consequences of anti-trafficking software tools.

5.1 Summary of Findings

This study provided information regarding the positive and negative consequences of anti-trafficking software through expert interviews. Anti-trafficking tools offer numerous advantages, including assisting law enforcement agencies in investigating trafficking cases and collecting information on criminal activity. Furthermore, these tools eliminate the need for manual analysis of large volumes of explicit data, saving time and resources. However, anti-trafficking tools also have negative consequences. They may infringe on privacy, endanger victims, enable traffickers to engage in adversarial efforts, negatively impact vulnerable populations, and result in negative social consequences owing to misclassification.

Participants explored strategies to reduce the adverse effects of anti-trafficking tools by examining their design and user considerations. The design aspects included working with relevant stakeholders, maintaining transparency, implementing privacy protection measures, and adopting an academic approach. User considerations focused on protecting the rights of sex workers, embracing data sharing, and preserving their privacy. These considerations aim to promote ethics and accountability in tool development.

5.2 RRI Implications

The AREA framework in RRI is designed to help researchers and innovators in various domains produce research results that are more closely aligned with societal needs and values. It plays a crucial role in ensuring that research and innovation projects are conducted in an ethical and

socially responsible manner. By integrating the principles of RRI into the design and implementation of a project, it is possible to engage with a variety of stakeholders in an inclusive and transparent way, thereby enhancing the acceptability of research or innovation by society at large [78].

The AREA framework can be employed at the beginning of the innovation cycle to identify mechanisms that can positively influence the trajectory of an ongoing project. In the field of computing, various elements of the AREA framework can be utilised to forecast research results at the beginning of a study. Although predicting future outcomes presents certain challenges, taking advantage of previous experience plays a critical role [56]. Researchers can adopt an approach similar to that demonstrated in this article and evaluate existing systems within their fields to gain a better understanding of the potential implications of their technology or research.

The AREA framework encourages researchers to engage in continuous reflection throughout the project design process. Although it is often used in the beginning of a project to determine whether to proceed with a research project [56, 62], it can also be applied to scrutinise various aspects of research, such as interview participants, findings, and methods. In doing so, researchers can uncover critical insights that could alter the trajectory of their projects. In this study, we employed this aspect of the framework when making decisions related to participant selection, to guide interview questions, and to identify a set of recommendations to develop and use anti-trafficking tools.

The AREA framework and previous literature [19, 62] recommend that researchers consult diverse groups of experts when examining the potential impacts of their research projects or innovations. Engaging with various groups of experts can lead to understanding of various issues that may not always be immediately apparent [78]. This is especially important in projects with social implications, as technical experts may not always fully understand the social consequences of their work; this is evident in our current work.

Researchers seeking to assess software in their projects can adopt a similar approach to this study, using the AREA framework to investigate the unintended consequences of software in their fields or projects. Incorporating quantitative methods into the research methodology can further enrich the data collected.

6 Conclusion and Future Work

This study aimed to assess the impact of Victim/Trafficker Identification tools through the application of a RRI approach and to provide recommendations to minimise any negative consequences that may arise from the use of these tools. The study used the RRI framework, AREA, to guide the data collection and analysis processes. Semi-structured interviews were conducted with 13 stakeholders including law enforcement officials, developers and anti-trafficking experts.

The research revealed three main themes: the positive and negative consequences of Victim/Trafficker Identification Tools and recommendations to mitigate the negative consequences. The adverse consequences revealed in this study offer a chance to improve the design and development processes of these tools, ensuring responsible results. The recommendations provided in this publication serve as an initial point for developing formal measures that anti-trafficking stakeholders can use in designing and employing anti-trafficking systems to mitigate negative outcomes.

Developing and implementing effective, responsible and beneficial anti-trafficking systems is essential for addressing the issue of human trafficking. Balancing the need to identify and address trafficking activities with the need to protect individuals from negative consequences is critical for stakeholders. Achieving this balance is important for ensuring that anti-trafficking efforts are successful in reducing the number of individuals affected by this crime.

This research is part of an ongoing study that seeks to develop a Victim/Trafficker Identification tool to identify trafficking activities online. In response to participants' advice, we shifted our

geographical focus from the US and the UK to 18 Sub-Saharan African countries, such as Kenya, Malawi, and South Africa, as these regions have limited tools and lower risks associated with deploying such tools. Future investigations will aim to confirm the applicability of our findings to the African context and expand on them through the use of both quantitative and qualitative research methods that adhere to the RRI approach. Using a mixed-method approach, we can gain a more comprehensive understanding of the research problem in Africa. Once we have built upon our findings, we will use these results to develop a formal framework for creating anti-trafficking software in Africa that can be adapted to other countries. Subsequently, we will employ this framework to develop an anti-trafficking tool as part of our project.

We urge researchers to also explore the following: (a) the impact of tools in the global south, as there is limited research on the effectiveness of anti-trafficking tools in this region, (b) extend the results to encompass a wider range of anti-trafficking tools, as we focus solely on a set of tools, such as awareness raising tools, which are the second most common tools used in anti-trafficking efforts [57], and (c) develop a standard set of measures for creating Victim/Trafficker Identification tools. This area has not been thoroughly explored, and stakeholders have called for more standardised measures [57].

References

[1] Wael Albayaydh and Ivan Flechais. 2023. *Examining Power Dynamics and User Privacy in Smart Technology Use Among Jordanian Households*. USENIX Association. Retrieved from https://www.usenix.org/conference/usenixsecurity23/presentation/albayaydh

[2] Pauline Anthonysamy, Awais Rashid, Ruzanna Chitchyan, and Security Lancaster. 2017. *Privacy Requirements: Present and Future*. Technical Report. Retrieved from https://www.researchgate.net/publication/318037140_Privacy_Requirements_Present_Future

[3] Hassan Marzoughi Ardakani. 2020. Identifying human trafficking networks in Louisiana by using authorship attribution and network modeling. *LSU Doctoral Dissertations* 5274 (2020).

[4] Renata Arronte. 2018. *A Partner in Trafficking: The Role of Internet Technologies in the Facilitation of Human Trafficking*. Ph. D. Dissertation. Utica College.

[5] T. Bai, J. Luo, J. Zhao, B. Wen, and Q. Wang. 2021. Recent advances in adversarial training for adversarial robustness. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*. Survey Track, 4312–4321. https://doi.org/10.24963/ijcai.2021/591

[6] Coryn Barclay. 2018. *Semi-Structured Interviews*. Technical Report. Retrieved from https://know.fife.scot/__data/assets/pdf_file/0028/177607/KnowHow-Semistructured-interviews.pdf

[7] David Barney. 2018. Trafficking technology: A look at different approaches to ending technology-facilitated human trafficking. *Pepperdine Law Review* 45, 4 (2018), 747.

[8] Christoph Bartneck, Christoph Lütge, Alan Wagner, and Sean Welsh. 2020. *An Introduction to Ethics in Robotics and AI*. Technical Report. Retrieved from https://library.oapen.org/handle/20.500.12657/41303

[9] BetterHealth. 2025. Stigma, Discrimination and Mental Illness - Better Health Channel. Retrieved January 1, 2025 from https://www.betterhealth.vic.gov.au/health/servicesandsupport/stigma-discrimination-and-mental-illness

[10] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. DOI: https://doi.org/10.1191/1478088706qp063oa

[11] Virginia Braun, Victoria Clarke, Nikki Hayfield, and Gareth Terry. 2019. Thematic analysis BT - Handbook of research methods in health social sciences. *Springer Nature Singapore Pte Ltd.* (2019), 843–860. DOI: https://doi.org/10.1007/978-981-10-5251-4_103

[12] Nadia Burkart and Marco F. Huber. 2021. *A Survey on the Explainability of Supervised Machine Learning*. Technical Report. 245–317 pages.

[13] C3.AI. [n. d.]. Feedback Loop. Retrieved February 12, 2024 from https://c3.ai/glossary/features/feedback-loop/

[14] Paolo Campana. 2022. *Online and Technology-facilitated Trafficking in Human Beings Summary and Recommendations Group of Experts on Action against Trafficking in Human Beings*. Technical Report.

[15] Lura Chamberlain. 2019. *FOSTA: A Hostile Law with a Human Cost*. Technical Report. 2171 pages.

[16] Christine Chen, Nicola Dell, and Franziska Roesner. 2019. Computer security and privacy in the interactions between victim service providers and human trafficking survivors. In *Proceedings of the 28th USENIX Security Symposium*. 89–104.

[17] Congress.Gov. 2022. S.3538 - 117th Congress (2021-2022): EARN IT Act of 2022 | Congress.gov | Library of Congress. Retrieved September 10, 2022 from https://www.congress.gov/bill/117th-congress/senate-bill/3538
[18] Julia Deeb-Swihart, Alex Endert, and Amy Bruckman. 2019. Understanding law enforcement strategies and needs for combating human trafficking. In *Proceedings of the Conference on Human Factors in Computing Systems*. Association for Computing Machinery. DOI: https://doi.org/10.1145/3290605.3300561
[19] 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. DOI: https://doi.org/10.1145/3555186
[20] Digital Regulation Cooperation Forum. 2022. Auditing Algorithms: The Existing Landscape, Role of Regulators and Future Outlook - GOV.UK. Retrieved November 17, 2022 from https://www.gov.uk/government/publications/findings-from-the-drcf-algorithmic-processing-workstream-spring-2022/auditing-algorithms-the-existing-landscape-role-of-regulators-and-future-outlook
[21] Niklas Donges. 2019. The Random Forest Algorithm: A Complete Guide | Built In. Retrieved June 16, 2021 from https://builtin.com/data-science/random-forest-algorithm
[22] Evelyn Douek. 2020. The Rise of Content Cartels | Knight First Amendment Institute. Retrieved September 10, 2022 from https://knightcolumbia.org/content/the-rise-of-content-cartels
[23] 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. DOI: https://doi.org/10.1080/23322705.2015.1015342
[24] Andrew C. Dwyer, Nathaniel O'Grady, Pip Thornton, Till Straube, Emily Gilbert, and Louise Amoore. 2021. Cloud ethics: Algorithms and the attributes of ourselves and others. *The AAG Review of Books* 9, 3 (2021), 36–49. DOI: https://doi.org/10.1080/2325548x.2021.1921458
[25] Erin Taylor. 2019. Sex workers are at the forefront of the fight against mass surveillance and Big Tech. The Observer (Nov. 12, 2019). Retrieved Feb 02, 2022 from https://observer.com/2019/11/sex-workers-mass-surveillance-big-tech/
[26] European Commission. 2019. Ethics Guidelines for Trustworthy AI | Shaping Europe's Digital Future. Retrieved January 10, 2025 from https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai?
[27] European Commission. 2020. Responsible Research and Innovation | Horizon 2020. Retrieved November 18, 2021 from https://ec.europa.eu/programmes/horizon2020/en/h2020-section/responsible-research-innovation
[28] Andrew Guthrie Ferguson. 2017. *The Smart Fourth Amendment, 102 Cornell L.* Technical Report. Retrieved from http://scholarship.law.cornell.edu/clrAvailableat:http://scholarship.law.cornell.edu/clr/vol102/iss3/1
[29] Erik Fisher. 2019. Technology assessment in practice and theory. *TATuP - Journal for Technology Assessment in Theory and Practice* 28, 2 (2019), 73–74. https://doi.org/10.14512/tatup.28.2.s73
[30] Inno Flores. 2022. Google AI Mistakenly Flags Parent for Child Sexual Abuse Material After Taking Photos of Toddler for Medical Use | Tech Times. Retrieved September 10, 2022 from https://www.techtimes.com/articles/279466/20220821/google-ai-flags-parent-child-sexual-abuse-material-medical-use.htm
[31] Luciano Floridi. 2016. On human dignity as a foundation for the right to privacy. *Philosophy & Technology* 29, 4 (2016), 307–312. https://doi.org/10.1007/s13347-016-0220-8
[32] Monica Franco-Santos and David Otley. 2018. Reviewing and theorizing the unintended consequences of performance management systems. *International Journal of Management Reviews* 20, 3 (2018), 696–730. DOI: https://doi.org/10.1111/ijmr.12183
[33] Alex A. Freitas. 2014. *Comprehensible Classification Models-a Position Paper*. Technical Report.
[34] Batya Friedman, Peter H. Kahn, and Alan Borning. 2022. *Value Sensitive Design: Theory and Methods*. Technical Report. https://research.cs.vt.edu/ns/cs5724papers/6.theoriesofuse.cwaandvsd.friedman.vsd.pdf
[35] Anne T. Gallagher. 2010. *The International Law of Human Trafficking*. Cambridge University Press.
[36] Felicity Gerry Q. C., Julia Muraszkiewicz, and Niovi Vavoula. 2016. The role of technology in the fight against human trafficking: Reflections on privacy and data protection concerns. *Computer Law and Security Review* 32, 2 (2016), 205–217. DOI: https://doi.org/10.1016/j.clsr.2015.12.015
[37] Ben Green. 2021. The contestation of tech ethics: A sociotechnical approach to technology ethics in practice. *Journal of Social Computing* 2, 3 (2021), 209–225. DOI: https://doi.org/10.23919/JSC.2021.0018
[38] Home Office. 2018. Modern Slavery Act 2015 - GOV.UK. Retrieved September 8, 2022 from https://www.gov.uk/government/collections/modern-slavery-bill#guidance-for-enforcement-officers
[39] Soraj Hongladarom. 2023. Shoshana Zuboff, The age of surveillance capitalism: The fight for a human future at the new frontier of power. *AI and SOCIETY* 38, 6 (2023), 2359–2361. DOI: https://doi.org/10.1007/s00146-020-01100-0
[40] Marisa Hultgren, Murray E. Jennex, John Persano, and Cezar Ornatowski. 2016. Using knowledge management to assist in identifying human sex trafficking. In *Proceedings of the Annual Hawaii International Conference on System Sciences*. 4344–4353. DOI: https://doi.org/10.1109/HICSS.2016.539

[41] Marisa Hultgren, Jessica Whitney, Murray E. Jennex, and Aaron Elkins. 2018. A knowledge management approach to identify victims of human sex trafficking. *Communications of the Association for Information Systems* 42, 1 (2018), 602–620. DOI: https://doi.org/10.17705/1CAIS.04223

[42] Marina Jirotka, Grimpe. Barbara, Bernd Stahl, Grace Eden, and Mark Hartswood. 2017. Innovation in the digital age. *China's Quest for Innovation* (2017), 195–215. DOI: https://doi.org/10.4324/9781351019743-10

[43] Justice and Home Affairs Committee. 2022. *Technology Rules? The Advent of New Technologies in the Justice System.* Technical Report. Retrieved from https://publications.parliament.uk/pa/ld5802/ldselect/ldjusthom/180/180.pdf

[44] Mark Latonero. 2011. Human trafficking online: The role of social networking sites and online classifieds. *Center on Communication Leadership & Policy*. University of Southern California, Los Angeles, CA, USA. September 2011. https://doi.org/10.2139/ssrn.2045851

[45] Mark Latonero. 2012. The rise of mobile and the diffusion of technology-facilitated trafficking. *SSRN Electronic Journal* November (2012). DOI: https://doi.org/10.2139/ssrn.2177556

[46] Mary Graw Leary. 2014. Fighting fire with fire: Technology in child sex trafficking. *Duke Journal of Gender Law & Policy* 21, 2 (2014), 289–323. Retrieved from https://scholarship.law.duke.edu/djglp/vol21/iss2/2

[47] Randy Lippert. 2008. David lyon, surveillance studies: An overview. *Canadian Journal of Sociology* 33, 2 (2008), 471–474. DOI: https://doi.org/10.29173/cjs2004

[48] Marinus Analytics. 2022. Traffic Jam – Marinus Analytics. Retrieved February 28, 2022 from https://www.marinusanalytics.com/traffic-jam

[49] Danae Metaxa and Jeff Hancock. 2022. *Using Algorithm Audits to Understand AI.* Technical Report.

[50] Kimberly J. Mitchell and Dana Boyd. 2014. Understanding the role of technology in the commercial sexual exploitation of children: The perspective of law enforcement. *Crimes Against Children Research Center* (2014). Retrieved from http://scholars.unh.edu/cgi/viewcontent.cgi?article=1036&context=ccrc

[51] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2023. Human-in-the-loop machine learning: A state-of-the-art. *Artificial Intelligence Review* 56, 4 (2023), 3005–3054. DOI: https://doi.org/10.1007/s10462-022-10246-w

[52] Towera Jessica Moyo, Omer Gunes, and Marina Denise Jirotka. 2025. Investigating human trafficking recruitment online: A study of fraudulent job offers on social media platforms. *Proceedings of the ACM on Human-Computer Interaction* 9, 2 (2025). DOI: https://doi.org/10.1145/3711016

[53] Alhassan Mumuni and Fuseini Mumuni. 2022. Data augmentation: A comprehensive survey of modern approaches. *Array* 16 (Dec. 2022), Article 100258. https://doi.org/10.1016/j.array.2022.100258

[54] Jennifer Musto, Mitali Thakor, and Borislav Gerasimov. 2020. Editorial: Between hope and hype: Critical evaluations of technology's role in anti-trafficking. *Anti-Trafficking Review* 14 (2020), 1–14. DOI: https://doi.org/10.14197/atr.201220141

[55] Ashley Olivine. [n. d.]. What Is Stigma? Examples, Impact, and Coping. Retrieved January 14, 2021 from https://www.verywellhealth.com/stigma-5215412

[56] ORBIT. 2021. The Keys of RRI | Orbit RRI. Retrieved November 18, 2021 from https://www.orbit-rri.org/resources/keys-of-rri

[57] OSCE. 2020. *Leveraging Innovation to Fight Trafficking in Human Beings: A Comprehensive Analysis of Technology Tools.*

[58] Richard Owen. 2014. The UK engineering and physical sciences research council's commitment to a framework for responsible innovation. *Journal of Responsible Innovation* 1, 1 (2014), 113–117. DOI: https://doi.org/10.1080/23299460.2014.882065

[59] Richard Owen, Jack Stilgoe, Phil Macnaghten, Mike Gorman, Erik Fisher, and Dave Guston. 2013. A framework for responsible innovation. *Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society* April (2013), 27–50. DOI: https://doi.org/10.1002/9781118551424.ch2

[60] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman. 2018. SoK: Security and privacy in machine learning. In *Proceedings of the 2018 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 1–19. https://doi.org/10.1109/EuroSP.2018.00035

[61] Melisa A. Pendergrass. 2018. The intersection of human trafficking and technology. Master's thesis, Utica College, Utica, NY. ProQuest Dissertations & Theses Global. Publication No. 10813486. https://www.proquest.com/openview/17f8d16abb6fb81a26f69dadebb19705

[62] Pili. 2017. Responsible Research Innovation. Retrieved November 18, 2021 from http://2017.igem.org/Team:Exeter/HP/Silver

[63] van de Ibo Poel and Lamber Royakkers. 2011. *Ethics, Technology, and Engineering.* Technical Report.

[64] Bhumika Ramani. [n. d.]. How to Ensure Explainability and Transparency of AI and ML Decisions. Retrieved February 09, 2024 from https://www.linkedin.com/advice/0/how-do-you-ensure-explainability-transparency

[65] Alfonso Ávila Robinson, Shuto Miyashita, and Shintaro Sengoku. 2023. *Characterizing ELSI/RRI/RI Frameworks and Their Links to Innovation Management Theory.* Technical Report.

[66] Aja Romano. 2018. FOSTA-SESTA, a Law Intended to Curb Sex Trafficking, Threatens the Internet's Future - Vox. Retrieved December 12, 2022 from https://www.vox.com/culture/2018/4/13/17172762/fosta-sesta-backpage-230-internet-freedom

[67] RRI Tools project consortium. 2017. RRI Tools: A project to foster responsible research and innovation for society, with society. *Funded by the European Commission under FP7 (Grant Agreement No. 612393)*. Retrieved from https://cordis.europa.eu/project/id/612393/reporting

[68] Siddhartha Sarkar. 2015. Use of technology in human trafficking networks and sexual exploitation: A cross-sectional multi-country study. *Transnational Social Review* 5, 1 (2015), 55–68. DOI: https://doi.org/10.1080/21931674.2014.991184

[69] Daniel Ribeiro Silva, Andrew Philpot, Abhishek Sundararajan, Nicole Marie Bryan, and Eduard Hovy. 2014. Data integration from open internet sources and network detection to combat underage sex trafficking. *ACM International Conference Proceeding Series* (2014), 86–90. DOI: https://doi.org/10.1145/2612733.2612746

[70] Judith Simon. 2016. Value-sensitive design and responsible research and innovation. In *Proceedings of the Ethics of Technology - Methods and Approaches*. London.

[71] Daniel J. Solove, Anita Allen, Howard Erichson, Jim Freeman, Robert Gellman, Ra-chel Godsil, Stan Karas, Orin Kerr, Raymond Ku, Chip Lupu, et al. 2006. *Formerly American Law Register A TAXONOMY OF PRIVACY*. Technical Report 477.

[72] Devin Soni. 2022. Feedback Loops in Machine Learning Systems | by Devin Soni | Towards Data Science. Retrieved February 12, 2024 from https://medium.com/data-science/feedback-loops-in-machine-learning-systems-701296c91787

[73] Spotlight. 2025. Spotlight: Human Trafficking Intelligence and Leads | Thorn. Retrieved August 29, 2022 from https://www.thorn.org/spotlight/

[74] Jack Stilgoe, Richard Owen, and Phil Macnaghten. 2013. Developing a framework for responsible innovation. *Research Policy* 42, 9 (2013), 1568–1580. DOI: https://doi.org/10.1016/j.respol.2013.05.008

[75] Lisa Rende Taylor and Mark Latonero. 2018. Updated guide to ethics & human rights in anti_trafficking: Ethical standards for working with migrant workers and trafficked persons in the digital age. Issara Institute. Retrieved from https://respect.international/updated-guide-to-ethics-and-human-rights-in-anti-trafficking-ethical-standards-for-working-with-migrant-workers-and-trafficked-persons-in-the-digital-age/

[76] Erin Tichenor. 2020. 'I've never been so exploited': The consequences of FOSTA-SESTA in Aotearoa New Zealand. *Anti-Trafficking Review* 2020, 14 (2020), 99–115. DOI: https://doi.org/10.14197/atr.201220147

[77] Edmund Tong, Cara Jones, Amir Zadeh, and Louis Philippe Morency. 2017. Combating human trafficking with deep multimodal models. *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* 1 (2017), 1547–1556. DOI: https://doi.org/10.18653/v1/P17-1142

[78] UK Research and Innovation. 2023. Research Integrity in Healthcare Technologies. Retrieved November 11, 2021 from https://www.ukri.org/councils/epsrc/guidance-for-applicants/what-to-include-in-your-proposal/health-technologies-impact-and-translation-toolkit/research-integrity-in-healthcare-technologies/responsible-research-and-innovation/

[79] UNESCO. 2021. *Recommendation on the Ethics of Artificial Intelligence*. Technical Report.

[80] United Nations. 2000. Protocol to Prevent, Suppress and Punish Trafficking in Persons Especially Women and Children, Supplementing the United Nations Convention against Transnational Organized Crime | OHCHR. Retrieved January 12, 2024 from https://www.ohchr.org/en/instruments-mechanisms/instruments/protocol-prevent-suppress-and-punish-trafficking-persons

[81] United Nations. 2004. *United Nations Convention against Transnational Organized Crime and the Protocols Thereto*. Technical Report. https://www.unodc.org/documents/middleeastandnorthafrica/organised-crime/UNITED_NATIONS_CONVENTION_AGAINST_TRANSNATIONAL_ORGANIZED_CRIME_AND_THE_PROTOCOLS_THERETO.pdf

[82] United Nations Office on Drugs and Crime. 2018. *Global Report on Trafficking in Persons*. Technical Report.

[83] U.S. Department of State. 2019. Trafficking in Persons Report 2019. Office to Monitor and Combat Trafficking in Persons, U.S. Department of State (June 2019). Retrieved from https://www.state.gov/reports/2019-trafficking-in-persons-report/

[84] UXPIN. 2024. What is UX Honeycomb and How it Impacts User Experience? Retrieved from https://www.uxpin.com/studio/blog/ux-honeycomb-definition-and-use/

[85] René Von Schomberg. 2013. *A Vision of Responsible Research and Innovation*. Wiley, DOI: https://doi.org/10.1002/9781118551424.ch3

[86] Jessica Whitney, Aaron Elkins, Murray E. Jennex, and Eric Frost. 2022. The use of emojis in online human sex trafficking ads. *Journal of Knowledge Management Practice* 23, 1 (Dec. 2022). https://doi.org/10.62477/jkmp.v23i1.3

[87] Till Winkler and Sarah Spiekermann. 2021. Twenty years of value sensitive design: A review of methodological practices in VSD projects. *Ethics and Information Technology* 23, 1 (2021), 17–21. DOI: https://doi.org/10.1007/s10676-018-9476-2

[88] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems* 135 (May 2022), 364–381. https://doi.org/10.1016/j.future.2022.05.014

[89] Doris Xin, Litian Ma, Jialin Liu, Stephen Macke, Shuchen Song, and Aditya Parameswaran. 2018. Learning efficiently over heterogeneous databases: Sampling and constraints to the rescue. In *Proceedings of the 2nd Workshop on Data Management for End-To-End Machine Learning, DEEM 2018 - In Conjunction with the 2018 ACM SIGMOD/PODS Conference*. Association for Computing Machinery, Inc. DOI: https://doi.org/10.1145/3209889.3209897

[90] E. Yaghmaei and I. R. van de Poel. 2020. *Assessment of Responsible Innovation*. Routledge, DOI: https://doi.org/10.4324/9780429298998

[91] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2021. A comprehensive survey on transfer learning. *Proceedings of the IEEE* 109, 1 (Jan 2021), 43–76. https://doi.org/10.1109/JPROC.2020.3004555

Received 13 March 2024; revised 31 January 2025; accepted 10 May 2025