Ethical Tensions in Applications of AI for Addressing Human Trafficking: A Human Rights Perspective

JULIA DEEB-SWIHART, Georgia Institute of Technology, USA
ALEX ENDERT, Georgia Institute of Technology, USA
AMY BRUCKAMN, Georgia Institute of Technology, USA

In the last two decades, human trafficking (where individuals are forcibly exploited for the profits of another) has seen increased attention from the artificial intelligence (AI) community. Clear focus on the ethical risks of this research is critical given that those risks are disproportionately born by already vulnerable populations. To understand and subsequently address these risks, we conducted a systematic literature review of computing research leveraging AI to combat human trafficking and apply a framework using principles from international human rights law to categorize ethical risks. This paper uncovers a number of ethical tensions including bias endemic in datasets, privacy risks stemming from data collection and reporting, and issues concerning potential misuse. We conclude by highlighting four suggestions for future research: broader use of participatory design; engaging with other forms of trafficking; developing best practices for harm prevention; and including transparent ethics disclosures in research. We find that there are significant gaps in what aspects of human trafficking researchers have focused on. Most research to date focuses on aiding criminal investigations in cases of sex trafficking, but more work is needed to support other anti-trafficking activities like supporting survivors, adequately address labor trafficking, and support more diverse survivor populations including transgender and nonbinary individuals.

CCS Concepts: • **Social and professional topics → Computing / technology policy**.

Additional Key Words and Phrases: ethics, human rights, artificial intelligence, human trafficking

ACM Reference Format:
Julia Deeb-Swihart, Alex Endert, and Amy Bruckamn. 2022. Ethical Tensions in Applications of AI for Addressing Human Trafficking: A Human Rights Perspective. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 295 (November 2022), 29 pages. https://doi.org/10.1145/3555186

1 INTRODUCTION

Over the last two decades, interest in applying artificial intelligence (AI) to combat human trafficking has increased – driven in part by increased awareness of the many ways in which human trafficking operations use online technology. This work incorporates methods like machine learning, computer vision, and social network analysis in a variety of anti-trafficking contexts including automatic victim identification[5, 48, 50, 51, 53–56, 76, 115, 146], evaluating policy outcomes[22, 109], identifying organization networks[99, 128], assisting with forensic investigations[9, 10, 28, 112, 134], and automatically detecting grooming behavior and child exploitation [37, 38, 103, 135, 140].

Authors' addresses: Julia Deeb-Swihart, jdeeb3@gatech.edu, jdeeb3@uw.edu, Georgia Institute of Technology, Atlanta, Georgia, 30332, USA; Alex Endert, endert@gatech.edu, Georgia Institute of Technology, Atlanta, Georgia, 30332, USA; Amy Bruckamn, asb@cc.gatech.edu, Georgia Institute of Technology, Atlanta, Georgia, 30332, USA.

2573-0142/2022/11-ART295
https://doi.org/10.1145/3555186

These approaches are increasingly being used in real investigations and are embedded in commercially available tools. For example, Amazon Rekognition, one of the most popular computer vision services, has been used by law enforcement since 2017 for cases relating to human trafficking [117]. This software was even credited as specifically instrumental in the 2019 arrest and prosecution of a trafficker [81]. While this work has the potential to save lives, prevent exploitation, and ultimately help end slavery worldwide, this research also faces a number of criticisms concerning potential dangers. A number of civil rights groups have noted major concerns that this work could further cause harm to marginalized communities, increase unwarranted governmental surveillance capacities, and violate privacy rights of citizens.

As these AI systems become more widely adopted and are therefore fundamentally embedded into our social systems, we must examine the implications of their development particularly with respect to their potential to cause harm. Additionally, more of the computing research community is starting to get involved in these efforts. In March of 2020, the Computing Community Consortium (CCC) hosted a workshop to create a research road map for efforts to apply AI to combating Human Trafficking. It is critical that we examine best practices as well as safeguards and strategies for harm prevention. By examining these efforts through the lens of human rights we can begin to uncover the ethical tensions associated with this research and highlight new avenues for future research centered on ethics. The human rights perspective uses the principles enshrined in international law by governing agencies like the UN, and this perspective affords a view that is centered on the individuals affected over technological advancements. Further, this framework uses a shared universal language which allows for cross-disciplinary understanding of our analysis.

The goal of this paper is to examine the current state of the art and understand the ethical tensions surrounding applications of AI in anti-trafficking research through a Systematic Literature Review. We use principles from human rights law as a lens to examine the prior work and guide our recommendations for future work. Additionally, this work represents a unique case study for examining the ethics of applied AI in domains that intersect the criminal justice systems and vulnerable populations.

This work presents an overview of the ethical issues present in applied AI research for human trafficking. We discuss these issues and gaps in work across eight human rights principles – Privacy, Accountability, Safety & Security, Transparency & Explainability, Fairness & Non-Discrimination, Human Control of Technology, Professional Responsibility, and Promotion of Human Values – and contextualize our analysis by drawing from prior work on ethics and AI. We conclude by presenting a series of "calls to action" for future research to address the ethical issues we found in our analysis. Our calls to action further offer insight as a case study on how to ethically apply AI to sensitive domains.

2 BACKGROUND AND DEFINITIONS

To better contextualize our work, we first define the term ”Artifical Intelligence" or AI so as to explicitly lay out the scope of this work. Then we give an overview of Human Rights-Based Approaches (sometimes referred to as HRAs) and situate the specific human rights framework we use for our analysis within the broader field of HRAs. Finally, we then briefly give an overview of human trafficking while highlighting the role technology plays in the ecosystemin order to describe the context in which these AI systems are designed to intervene.

2.1 Defining AI

Defining Artificial Intelligence (AI) has long been a contentious issue among researchers and practitioners [39, 73]. The field of AI is evolving at an unprecedented rate and prior definitions have not kept up-to-date with modern conceptualizations of AI. To this day there remains no consensus on a universal definition of AI [73].

To accommodate the many possible conceptualizations of AI, we take a broad and descriptive definition of AI that encompasses a wide range of sub-fields including machine learning, visual analytics, and data science. Papers included in our analysis are not limited to those that are explicitly labeled AI, but instead involved either A) some sort of machine-assisted processing of data and/or B) the creation of a computing system that assists with decision making. Our goal with taking such a broad view of AI was to ensure that our work included a wide range of methods and approaches. Our definition of AI is heavily inspired by definitions listed in AI governance policy documents from AccessNow [11] and the European Commission's High Level AI Expert Group [49] and only excludes methods that are exclusively qualitative or exclusively hardware oriented.

2.2 Overview of Human Rights-Based Approaches (HRAs)

Human-Rights based approaches use the values and principles that underlie international law to guide and inform development; evaluate current research efforts; and guide public policy discussions [68]. These principles are derived from internationally recognized legal frameworks, such as the United Nation's Declaration of Human Rights (UNDHR) [13], and include fundamental principles like dignity, fairness, liberty, and equality. As a framework, human-rights approaches take a people-centered approach to analysis by drawing attention to the rights of those affected by development and the responsibility for researchers to examine practices for preserving human rights.

In our paper, we contextualize our analysis using the 8 human rights principles – *Privacy, Accountability, Safety & Security, Transparency & Explainability, Fairness & Non-Discrimination, Human Control of Technology, Professional Responsibility, and Promotion of Human Values* – outlined by legal scholars in [39] and described in the list below. This work derived its principles by examining shared themes found in over 20 documents for AI governance put forth by civil society, governments, inter-governmental organizations, and the private sector. As this framework included a broad geographic representation not centered specifically on American or European laws, these 8 principles represent an international perspective to rights-based approaches for AI development.

- *Privacy* - AI systems should respect a person's right to privacy both in the data collection process and in the design and deployment of an AI system. People have the right to control the use of data about them and AI systems should respect their agency in how their data is used.
- *Accountability* - This principle is concerned with mechanisms for determining responsibility for the AI systems and ensuring said responsibility is appropriately distributed. Researchers must consider the impacts of an AI system and provide adequate remedies for those impacts.
- *Safety & Security* - AI systems need to reliably perform as expected without causing harm and not be vulnerable to attacks.
- *Transparency & Explainability* - Systems need to be designed for oversight through accessible and transparent design and through human understandable outputs.
- *Fairness & Non-discrimination* - Systems must be designed to prevent algorithmic bias and discrimination such as through developing representative datasets and through designing systems to be inclusive.
- *Human Control of Technology* - AI systems in critical decision making contexts must be subject to human review.
- *Professional Responsibility* - Researchers must ethically design AI systems and consider the long-term consequences of their designs. This principle emphasizes the values of accuracy, responsible design, multi-stakeholder collaboration, and scientific integrity.

- *Promotion of Human Values* - Systems should reflect core values and promote human well-being. AI should used to benefit society and not to promote harmful practices, discrimination, and unequal conditions.

2.3 Defining Human Trafficking

Human trafficking is a complex crime that takes a variety of forms. The circumstances and situations each trafficking victim faces varies drastically from person to person; thus, historically, experts have disagreed upon a universal definition of human trafficking. However, the most commonly cited definitions of human trafficking (and therefore the definitions we use within this paper) are established by two pieces of legislation: the US Trafficking Victims Protection Act of 2000 (TVPA) [2] and the UN's Palermo Protocol of 2000 [130].

Under both these laws, human trafficking refers to the crime of recruiting, harboring, transporting or obtaining a person for the purposes of exploitation through the use of force, fraud, and/or coercion [2, 130]. Human trafficking is often described across two different categories of trafficking: labor trafficking and sex trafficking. Labor trafficking refers to cases where a person is forced against their will to work in an otherwise lawful industry such as agriculture or domestic work [133]. Labor traffickers often lure victims through false promises of gainful employment and then force them to work long hours for little to no pay in unsafe working environments [133]. Most commonly, victims of labor trafficking are enslaved through a practice called debt bondage or debt slavery, where a person is forced to work to pay off a personal debt [58]. Sex trafficking, on the other hand, refers to cases where a victim is forced to perform commercial sex acts such as prostitution or pornography for the financial gain of others [133]. These two categories are not mutually exclusive and there have been cases where a person is a victim of both labor and sex trafficking [1].

the International Labor Organization (ILO) estimates that 89 million people have experienced some form of trafficking in the last 5 years[58]. Currently there is estimated to be over 40.3 million victims of human trafficking – of which 24.9 million are estimated to be victims of labor trafficking and 4.8 million are estimated to be victims of sex trafficking [58]. Of these, one in four victims of human trafficking are children – including over 1 million children who are victims of sex trafficking [58].

2.4 Human Trafficking and Technology

Increasingly internet technology has become an integral part of human trafficking operations [19, 74]. Both the UN and the US now report that the majority of human trafficking court cases have involved some use of internet technology [19, 132]. Traffickers have used social media and other online sites to identify potential victims [74, 132]; recruit and groom victims [90]; advertise and sell their victims [74, 90, 132]; launder and transfer money between trafficking organizations [99, 132]; and even to surveil their current victims [132]. In particular, social media sites are increasingly being used by traffickers to recruit victims – roughly half of the cases reported to the UN involved traffickers recruiting victims through social media [132].

Traffickers have also used internet technologies to directly exploit their victims, leading to a new form of trafficking sometimes called "cyber-trafficking." Cyber-trafficking is a subset of sex trafficking that involves a trafficker forcing victims to perform sex acts that are streamed online [26]. For example, one trafficking organization in the Philippines operated a "cybersex den" where victims were kidnapped and forced to perform sex acts for online customers through live-streaming

[1] A recent example involving both labor and sex trafficking can be seen in the "cantina cases" documented by the National Human Trafficking Hotline where victims were forced by their trafficker to both work in bars as unpaid waitresses and provide commercial sexual services to the bar customers [33].

services [1]. This type of trafficking is relatively new as it has only become possible by widespread access to high speed internet, cheaper high quality webcams, and the widespread adoption of live-streaming social media sites [26].

While these online sites facilitate human trafficking, they also provide vital information for identifying and tracking cases of human trafficking. Online advertisements and social media posts leave behind digital traces that allow law enforcement to track these activities and have played a crucial forensic role in many criminal investigations [19, 74]. For computing researchers focusing on combating human trafficking, these online traces provide crucial a source for datasets. Many of the computational models used to identify victims online were trained on datasets gathered from online sites where trafficking occured [5, 48, 50, 51, 53–56, 76, 115, 146].

3 RELATED WORK

There has been a long history examining the ethics of computing research across multiple domains, with more recent efforts focusing on applications of AI [44]. Prior work has, for example, examined the ethical tensions that arise from using large-scale social media datasets to train computational models [20, 24, 27, 147], examined how AI systems embed existing discriminatory values and biases [14, 93], and analyzed the human impact of large scale data collection such as issues of data privacy and algorithmic surveillance [20, 59, 84]. Others, such as the work in [70, 144, 144], further examine the role of AI within criminal justice settings and note that AI systems are fundamentally changing decision making processes.

Alongside these efforts, scholars have also looked to develop principles and frameworks to use as a critical lens to tease apart the ethical tensions present in AI research. In particular, there has been an increased attention on using the language of human rights as a framework to guide AI development. Thus far much of this work has focused at the macro-level – examining governing and policy practice for AI development [39] – but recently authors have called for using this framework to guide academic work as well [3, 68, 75, 116].

Several authors have noted that using human rights has several advantages as a framework over traditional ethics-based frameworks [68, 75, 83]. The human rights framework is based on internationally recognized legal standards and thus affords a shared, universal understanding. The principles that underlie this framework tend to be more concrete and center the analysis on the rights of the individuals affected by technology [75, 116]. This in turn forces the analysis to focus on people first over technological considerations – much in the same way as Value Sensitive Design and Participatory Design [3, 68]. Additionally, there is a natural overlap between human rights and other ethics-based frameworks [68]. The principles [39] we use in our analysis includes the 3 represented by the FAT model (Fairness, Accountability, Transparency). In practice, rights-based approaches have been used by scholars and policy experts to recommend best practices for preserving data privacy during pandemic response [82], highlight ethical considerations when using big data in healthcare settings [107], and to access harm cause by algorithms used in key decision-making contexts [83].

However, human-rights based approaches are not without their limitations. Using human rights as a lens tends to focus more on a macro-level and thus can miss nuance found at smaller scales [129]. The language of human rights is taken from legal documents and public policy briefs and thus presents barriers to participation and understanding especially outside the legal field [97]. Human-rights-based policies often feature unclear enforcement mechanisms for systems that span multiple states and involve non-state participants – a scenario found more often than not in AI development [97]. Human-rights-based policies further tend to feature a heavy emphasis on legal mechanisms to combat inequality and discrimination – which in some circumstances has a positive effect, but in others can secure political power for certain groups at the expense of others [23]. Finally, many of the values – including those used in this paper – heavily feature industry perspectives including from organizations with histories of problematic AI research. Many researchers and policy makers have expressed skepticism on principles derived and implemented by industry participants who may be motivated to incorporate human rights and other ethics-based principles for the purposes of "ethics washing" their brand [40, 141]. Thus, prior works has suggested that applications of human rights be careful not to de-emphasize the principles of accountability, transparency, and participation [40] .

Finally, it should be noted that human rights and ethics are complementary frameworks. There is a real benefit to synthesizing these perspectives. To this end, while we explicitly highlight a human-rights lens in our analysis, our discussion and interpretation of these principles is informed by our experiences with ethics frameworks. We particularly tried to emphasize values of equal participation and just design.

4 METHODS

Our work performs a Systemic Literature Review (SLR) using the methods outlined in [95] and [18]. At a high level, this process involved developing a set of keywords to search for papers, gathering the manuscripts captured by those terms from computing related publication databases, systematically pruning the resulting corpus to remove irrelevant material, and then analyzing the final corpus using a systematic data extraction and synthesis process.

4.1 Search and Screening Process

To construct our corpus, we conducted a literature search using a set of human trafficking related keywords to identify potentially relevant work. Because our the primary focus of our search was to identify work specifically at the intersection of AI and Human Trafficking, our search process targeted 3 main computer science publication databases(IEEExplore[2], The ACM Digital Library[3], Springer- Link[4]), and we used the following keywords and their derivatives to perform the queries: "human trafficking", "sex trafficking", "labor trafficking", "modern slavery", "sexual exploitation", "labor exploitation", "forced prostitution", "forced labor", "forced marriage", "trafficking survivors". Finally, we limited our search to only English peer-reviewed results that were published after 2000. We impose this time restriction to ensure a consistent definition of human trafficking; note that 2000 is the year when the US passed the TVPA and United Nations passed the UN Trafficking in Persons Protocol (Palermo Protocol). In total, 616 unique results were collected - 213 from the ACM Digital Library, 316 from IEEE Xplore, and 87 from Springer-Link.

We then constructed a practical screen using the criteria outlined in [95] to ensure that our final corpus only contained relevant material. Our screen took the form of a check-list and one of the authors was tasked with systematically classifying the articles as relevant or irrelevant to the survey. Articles are excluded if the answer to any of the questions below was no, otherwise the article was included in the final corpus:

(1) Does the study have a clearly stated or heavily implied application for human trafficking?
(2) Does the study have methods which involved either fully automated or machine-assisted processing of data? OR Does the study involve the creation or design of some computing tool?

[2] https://ieeexplore.ieee.org/Xplore/home.jsp
[3] https://dl.acm.org/
[4] https://link.springer.com/

Our inclusion and exclusion criteria had the effect of excluding papers whose methods are exclusively qualitative (such as ethnographic studies) and papers exclusively concerning hardware design. Because our emphasis was on examining AI practices, the papers who used these methods fell outside of the scope of our study. We did however include studies that used mixed-methods approaches as long as those studies involved the design or implementation of an AI system. Examples of papers that were excluded because they did not fit our definition of AI included interview studies to understand the socio-technical needs of anti-trafficking non-profits [118, 119], sex worker charities [120], law enforcement working human trafficking cases [35], and Nepalese survivors of sex trafficking [41]. However while these papers were excluded, the analysis and lessons learned from this work helped inform and shape our analysis. In particular, we found that many of these studies had valuable insights for how future research could be developed in a way that centers the needs of marginalized groups and we include these insights within our discussion. We also excluded papers which only incidentally matched our keywords which included papers discussing the “master-slave" computer architecture and papers focusing on automotive accidents. This results in a final corpus of 69 papers.

As part of the screening process, the author also noted if any of the papers were developed by the same team of authors concerning the same AI system. Our goal was to group papers written about the same system and treat those papers as one unit for analysis. However, we found that while two papers in our corpus were written about the same system, these papers described discrete and separate components of a larger system. After careful consideration, we decided to not group these two papers because each of the papers were substantially different enough to consider each of these studies their own AI systems.

4.2 Data Extraction and Synthesis

Our process involved two stages of analysis: 1) data extraction and 2) qualitative analysis. We use the results of the data extraction process to summarize the current state of the art and uncover gaps in the existing literature. We use the results of the qualitative analysis to form the taxonomy of ethical tensions described in the discussion section.

For data extraction, we use methods described in [69] and [95] to create a data extraction form to analyze the texts. This form contained qualitative criteria including: 1) Who is the intended user? 2)What is the intended use-case? 3) What type of human trafficking does this paper address? 4) What methods did the paper use? 5) What data did the paper use and where did the data come from? The form also tracked meta-information about the publication including the year of publication, the publication venue, the location of the study, and the institutions of the authors. Once the form was created, one of the authors performed the data extraction and used the "test-retest" process [69] to ensure consistency. The "test-retest" process involves the researcher re-extracting data from a random sample of the texts in the corpus and checking for consistency in answers to the form.

Finally, we performed a meta-analysis of the corpus using framework synthesis [18]. Framework synthesis takes a structured approach to qualitative data analysis, and uses an iterative, deductive process where codes are generated against an chosen framework [18]. For our analysis, we use the 8 human rights principles – *Privacy, Accountability, Safety & Security, Transparency & Explainability, Fairness & Non-Discrimination, Human Control of Technology, Professional Responsibility, and Promotion of Human Values* – described in [39] as our framework. The authors iteratively generated codes from the text based on relevance and/or absence of discussion relating to 8 human rights principles. As mentioned earlier, the goal of this analysis was to uncover themes around the ethical tensions noted in the papers as well as tensions not present in the papers. During this synthesis process, the authors also took detailed notes about the any ethical concerns or areas where future research is warranted. Results from the framework synthesis process were then used to narrative describe the existing research space and the structure the discussion of the ethical tensions.

4.3 Limitations

We note the following limitations in our methods choices. First due to practical considerations, our approach only uses one reviewer to judge the inclusion/exclusion of papers in the final corpus. While other papers utilize the same approach [36], using only 1 reviewer does increase the risk of bias and human error affecting the results our analysis.

The range of search criteria (both keywords and time frame) was limited to maintain a manageable set of papers to consider for the final corpus. Thus, some relevant papers may have been excluded that did not meet this criteria. In addition, because we limited our keywords to only those directly referencing human trafficking, we may have missed papers that might be relevant to human trafficking but do not explicitly mention human trafficking within the text of the publication. Finally, because we are only considering academic work, we must consider that some applications of AI for human trafficking exist outside of academic work (for example, in commercially available tools) and are therefore excluded from our analysis.

Finally our analysis uses human-rights as framework to uncover ethical tensions. Thus, our analysis might miss ethical tensions that don't align well with any of the principles. Alternative frameworks such as transformative justice could provide different views on this issue.

5 FINDINGS

We note two major trends in publication timing: 1) there was a surge in publications post 2016 and 2) that the popularity of this topic is continuing. At the same time, this remains an understudied issue. Our corpus found only 69 papers published in the last two decades. While human trafficking has gained more awareness over the years, this field remains relatively niche.

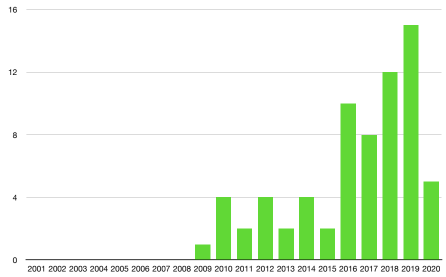

Fig. 1. Histogram of publications per year, 2020 includes partial data

5.1 Problem Areas & Use Cases

We identified five sub-domains of human trafficking that the papers address:

(1) *Labor Trafficking* - covers papers exclusively concerning forced labor, labor exploitation, and/or labor trafficking.
(2) *Sex Trafficking* - covers papers exclusively concerning forced prostitution, sex trafficking, and/or sexual exploitation but was not specific to only child cases.
(3) *Commercial Sexual Exploitation of Children (CSEC), Child Trafficking, and/or Child Pornography* - This category is shortened in Table 1 as "CSEC" and covers papers targeting victims of

Table 1. Table showing the breakdown of papers by intended use case and targeted area of intervention. Categories in this table are not exclusive - meaning papers can be assigned to multiple categories (ex: papers that address both labor and sex trafficking) and can be assigned to multiple use cases (ex: papers that target both aiding criminal investigations and inform policy). The number in parenthesis across the top represents the total number of papers that were as labor, sex, ect.

	Labor (3)	Sex (28)	CSEC (26)	Broadly HT (5)	One of Many (7)	Total
Aid Criminal Investigation	1	25	19	1	10	56
Inform Policy	2	2		2	2	8
Prevention		1	8	1		10
Support Survivors				1		1
Unclear					1	1

sex trafficking that are minors. Note that in our corpus, there were no papers targeting specifically child labor trafficking.

(4) *Broadly Human Trafficking* - covers papers whose applications either specifically targeted cases of both labor and sex trafficking or whose methods are intended to apply in circumstances where both labels would apply (such as Cantina workers who are both forced to provide labor for the restaurant and provide sexual services for the customers).

(5) *Human Trafficking as one of many possible use-cases* - covers papers where human trafficking was listed as one of many possible use cases. Many of these papers were designed to target other crimes that occur alongside human trafficking such as gang violence, organized crime, and drug trafficking. This category is shortened in Table 1 as "One of Many."

Results for the distribution of papers across these sub-domains can be seen in Figure 2. Combined the sex trafficking and CSEC categories represent roughly 78% of papers in the corpus - meaning that overwhelming majority of papers targeted sex trafficking more broadly. Only 3 papers listed labor trafficking as the primary focus.

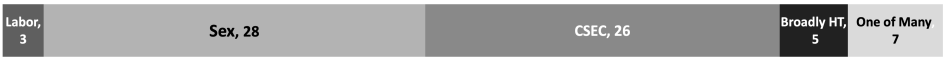

Fig. 2. Bar chart showing the distribution of papers in the corpus across the 5 problem areas. Combined, sex trafficking of adults and minors represents the most common focus of papers in the corpus

We also analyzed what the intended use case was for the results and tools developed in the paper. At a high level, these use-cases fit into the following categories (Note that we also included the category of "unclear" to capture papers who had unclear use-cases):

(1) *Aiding Criminal Investigations*
(2) *Informing Policy Decisions*
(3) *Preventing Human Trafficking and Educating the Public*
(4) *Supporting Survivors of Human Trafficking*
(5) *Unclear*

The categories we describe above broadly line up with the "4P Paradigm" - a framework established in both the TVPA and Palermo Protocol that categorizes anti-trafficking efforts [94]. This paradigm classifies anti-trafficking activities into 4 categories: prevention, protection, prosecution, and partnership [94, 102]. Prevention refers to activities centered on educating the general public, identifying vulnerable populations, and providing services and vocational alternatives for vulnerable groups. Protection refers to actives centered on rescuing, rehabilitating, and reintegrating survivors of human trafficking. Prosecution refers to activities centered around creating and enforcing anti-trafficking laws. This category also includes efforts to support existing criminal justice practices. Finally, partnership refers to activities centered on identifying and fostering collaboration and information sharing between anti-trafficking groups.

The 4p paradigm is used to evaluate anti-trafficking efforts and highlight areas where more work is needed [71]. We can use this model to see high-level patterns in what researchers have focused on and where future research is needed. Thus, in the sections below we note the overlaps between our findings and the categories in the paradigm.

5.1.1 Aiding Criminal Investigations. This category represents by far the largest use case with roughly 75% of papers (56 total out of 75) explicitly indicating this use case. Papers in this group predominantly dealt with applications of machine learning for automatic victim identification [5, 48, 50, 51, 53–56, 76, 115, 146] including specifically identifying child pornography [140], characterizing human trafficking activity [43, 76, 124], and detecting illicit behavior [38, 46, 127] or organizations [99, 128]. Using the 4P paradigm, the papers in this category broadly address "prosecution" as these papers were intended to support criminal justice efforts.

The most common task within this category was victim identification either through developing computational models to detect trafficking online or through developing public reporting systems [105, 123]. Research focusing on computational approaches for automatic victim identification use online sites like social media, dark web forums, and sex work ad sites to train models to detect cases of human trafficking. The goal of this work is to narrow down possible leads for law enforcement by differentiating posts about human trafficking from posts about other topics. There are a variety of approaches used for this task, though most borrow methods from Natural Language Processing (NLP) and machine learning. Many of these approaches rely on expert generated keywords to train models [50, 140]. For example, [56, 146] used supervised learning models trained on keywords and [5, 6, 48] used a mix of keywords and expert labels to both train and evaluate semi-supervised learning models. Beyond keywords and text-based approaches for victim identification, other work in this corpus used metadata present in the posts, images present alongside posts, and even payment data [99] to detect instances of human trafficking. For example, [53–56] used the location metadata present in Backpage[5] posts alongside keywords to detect instances of sex trafficking. Both [46, 127] use a mixture of NLP and computer vision to detect cases of child sex trafficking and use Convolutional Neural Networks (CNNs)[6] to estimate age and gender in images accompanying suspicious social media posts. Similarly, [65, 66, 115, 139] all use a mix of textual based approaches (like keywords), computer vision approaches, and metadata analysis to detect sex trafficking in online sites. [121, 122] use computer vision to identify distinguishing visual features present in the background of images posted alongside advertisements to geolocate the victims by matching those features to photos of hotel rooms.

Within the broader category of “Aiding Criminal Investigations” there were also papers that focused on the task of detecting trafficking organizations/networks. Many of these papers used similar data sources and methods as the research focused on the task of victim identification. For example, [76] used unsupervised template matching to uncover human trafficking organizations

[5] Backpage was an online classified site similar to Craigslist that included sections where people could advertise adult services and solicit sex work. However, traffickers have also used this site to advertise their victims [19, 74]. As a result in 2018, the website was seized by the US Department of Justice and was shut down.

[6] CNNs is a type of deep neural network used for image classification. This method falls broadly under the area of "Deep Learning"

and connections between advertisements posted to Backpage. Further, [99] used patterns in bitcoin payment information to uncover human trafficking organizations. Many of the approaches used various social network analysis (SNAs), Natural Language Processing (NLP) and community detection methods to uncover the groups [12, 25, 43, 51, 63, 64, 128] . The goal of this work is to support law enforcement investigations by allow investigators to uncover connected profiles based on similarities, interactions, and relationships with identified instances of trafficking.

Other papers dealt specifically with identifying grooming behaviors and child exploitation online. These papers used NLP and machine learning methods such as formal concept analysis[37], and various supervised learning approaches [38, 103, 135, 140] to categorize grooming behaviors in chat logs. The goal of this work is to assist law enforcement with assessing threat levels related to sexual abuse of minors.

Other papers in this category related to applications of computer vision models for forensics - including matching identifiable features in sexual abuse imagery [9, 10, 28, 112, 134] and tracking unique characteristics present in other human trafficking related images [121, 122]. Papers [15, 112, 134] all tackled the problem of vein pattern visualization where computer vision techniques are used to locate vein patterns seen on body parts to use for identification purposes. This application is particularly important for identifying the perpetrators in child pornography as faces tend to be obscured [112]. Additionally, to aid in the identification of child pornography and sexual abuse imagery, papers used various deep-learning models to estimate ages [8–10] and assist with generating age progression photographs [28].

Finally, this category also included papers that designed tools to support law enforcement investigations through building tools to organize leads [78, 98, 113], visualizing case data [4, 66], and improve search capabilities[50, 61, 96, 100, 110, 139, 145].

5.1.2 *Informing Policy Decisions.*

Papers in this category used computational approaches to evaluate current policy and programs. Many papers borrowed methods from operations research and used simulations to evaluate the efficacy of current policies. For example, [22] used system dynamics simulation models[7] to evaluate outcomes based on different labor policy implementations and [109] similarly used system dynamics models but focused on policies targeting sex trafficking. Much of this work relies on data from national reports. For example, [45] constructed transnational flow models using both quantitative and qualitative data within the US State department's annual Trafficking in Persons (TIP) report.

Others like [34] focused on improving risk assessments that identify vulnerable populations. Their goal was to better inform policy and governmental practices for preventing instances of trafficking. Additionally, [124] sought to evaluate the impact of environmental disasters (like hurricanes) on human trafficking patterns and whether these event increase the risk of victimization within already vulnerable populations. The goal is to both better understand the human trafficking ecosystem and to link the potential relationship between climate change and the increased risk for victimization.

Finally, papers in this category included [77, 111] which sought to use machine learning to improve prevalence estimation. For example, [77] specifically targets labor trafficking and uses time series analysis to estimate the scale of human trafficking in particular regions in India. Prevalence estimation is important for policy makers as these estimates provide context for decision-making and resource allocation.

Papers within this use case could align across all the categories in the 4P paradigm depending on the use case. Much of the work for this use case emphasizes evaluating current policy practices or estimating current population to better inform policy decisions, while placing less of an emphasis

[7] System dynamics simulations use models to represent cause-and-effect relationships and equations to simulate system behavior. Example models include causal-loop diagrams, and flow diagrams [108]

on policing or establishing laws. Thus, while some papers mention how their work could improve criminal justice policy and prevent trafficking, these papers perhaps better fit somewhere between prosecution and prevention.

5.1.3 Preventing Human Trafficking and Educating the Public. As the name of this category implies, papers in this category focused on preventing trafficking from occurring and explicitly fell under "Prevention" within the 4P paradigm. Papers that focused on preventing cases of trafficking tended to focus on either developing computational models to predict high risk individuals [72, 101] or on developing models to detect grooming behavior[8] so that interventions can occur sooner [37, 85–87]. This is distinct from the risk assessments described in the section above because the papers in this category are specifically intended to improve online moderation rather than inform governmental policy decision makers. Further, these predictions are intended to only be used in online settings and cannot be used to predict risk of victimization outside of an online context.

Additionally, some of the papers in this category further focused on developing systems to educate the general public about internet safety and human trafficking indicators. For example, [88] built a video game to teach children about harms online.

With the exception of [72], all of the papers in this category exclusively focused on preventing the trafficking of children and none focused on preventing labor trafficking.

5.1.4 Supporting Survivors of Human Trafficking. Only one paper fell within this category – [42] built a web-application to support Nepalese survivors needs for peer-bonding and social support. This work corresponds with "Protection" within the 4P paradigm.

5.1.5 Unclear. We labeled one paper, [31], as unclear. The authors of [31] built the tool, “Cyber Trafficking Surveillance System (CyTrass)”, to analyze discussions of cyber-trafficking related discussion on social media. However, it was unclear how the authors intended this system and its insights to be used in practice.

5.2 Intended Users

The overwhelming majority of papers (73%) explicitly list law enforcement as an intended user of the system. Interestingly, very few of these papers mentioned engaging with law enforcement as part of the design and development process despite the focus on law enforcement applications.

Other users mentioned included governments, NGOs & non-profit organizations, policy makers, military, the general public and parents of young children.

5.3 Discussion of Ethics

In our analysis, we also analyzed the discussions of ethics and limitations within the paper’s text (if present). We took a broad definition of ethics for this analysis which included any discussion of impacts and potential harms and any discussion that considered ethics principles in the research process.

Only 11 papers had explicit mentions of ethics either as a distinct section or embedded within another section. Within these papers, most focused primarily on issues of fairness and bias stemming from datasets. For example, papers [8, 10, 28] assessed the accuracy and bias present in facial recognition software used by law enforcement. These papers all note that many of the datasets that underlie commercially available computer vision software are unbalanced with respect to race, gender, and age; The result is that many of these tools have lower accuracy on faces that are young, feminine, and/or have darker skin [28, 112].

[8] grooming is a term used to describe the manipulative tactics adult sex offenders use to form connections with children for the purposes of future victimization

Additionally, only 2 papers explicitly mentioned using an ethics framework to guide their analysis. [57] examined the privacy implications of image classification used to detect child pornography and used the "Personal Information Protection and Electronic Documents Act" or PIPEDA principles[9] to guide their analysis and research process. Using principles from prior work on bias and AI systems, [51] proposed a bias mitigation plan as an inherent part of their research approach and tool design. Their plan included steps to both diagnose potential sources of bias and steps to mitigate those biases.

To some extent, the overall lack of ethical discussions can be explained by paper length limitations and the differing conventions for paper formats at different computing conferences. While limitation sections are common, distinct ethics sections are not.

6 APPLICATION OF THE HUMAN RIGHTS FRAMEWORK

In this section, we use the human rights principles established in [39] to analyze our data and provide a series of recommendations for future work. This paper seeks to answer what it means to address these 8 principles – Privacy, Accountability, Safety & Security, Transparency & Explainability, Fairness & Non-Discrimination, Human Control of Technology, Professional Responsibility, and Promotion of Human Values – with respect to the development of AI for anti-trafficking. We use this lens because it provides a universal vocabulary, well-developed standards and principles, and a concrete framework for solutions. Note that due to the high degree of overlap in our analysis, we condensed the principles "Privacy" and "Safety & Security" into one section.

6.1 Privacy, Safety, & Security

Privacy, Safety and Security have long been established in human rights law, and privacy is considered a fundamental human right under the UN's Declaration of Human Rights (UNDHR) [13]. In the context of AI research, these principles are concerned with the idea that AI systems should respect an individual's right to privacy – particularity with respect to data privacy, and that AI systems and their underlying datasets should be secure and resistant as much as possible to compromise.

The privacy and security concerns we found in our analysis using these principles stemmed from the datasets that underlie the models – raising issues around the storage, collection, and access to data as well as concerns about the scope of data collection and its impact on marginalized groups. In this section, we will first focus on issues of data collection scope and impact, and then focus on issues relating to long-term data stewardship and security practices. We noted three themes in the gaps in how papers addressed privacy. First through their dataset collection process and tool designs, we noted that several papers advocated either directly or indirectly for mass-scale surveillance and are thus potentially violating peoples' right to privacy. Second, we noted instances of researchers potentially exposing identifiable data either within the paper itself or through the release of non-anonymized public datasets. Finally, we find that this research raises questions about data access and protection.

We found a number of papers whose work – either directly or indirectly – advocated for mass-scale surveillance and used crime prevention as justification. This is a concerning theme to find because some cases of human trafficking have been linked to corruption and exploitation by user groups identified in the corpus [92, 123]. Additionally, surveillance for the purposes of crime prevention historically has also been used for human rights abuses and remains a contentious civil rights debate [7, 30, 60, 131]. While this work represents avenues for social good, we have to examine what limitations should be considered for collecting data and what impact this work will

[9] PIPEDA is federal data privacy legislation in Canada

have on different groups. These privacy concerns were especially prominent in models designed to detect instances of human trafficking in online activity. For example, [86] developed a tool for parents that monitors a child's online activity and uses AI to detect instances of grooming behavior. Similarly, [101] developed a system of connected IoT tools that tracks children's movement patterns via GPS to prevent cases of child abduction for the purposes of sexual abuse. While human rights law is murky concerning children's rights to privacy, many of these applications – through the collection and storage of data concerning these children's activities – create potential avenues for harm. This work often ignores the potential for adults to use these systems for abuse and control and also ignores the very real cases where human trafficking is facilitated by a parent or close relative. Based on reports to service providers, between 10 and 30% of human trafficking cases are facilitated by family member [92]. Further from a security perspective, the collection of data on children's activities if compromised might further put vulnerable children at risk for exploitation. Recent cases like IoT baby monitors being hacked by pedophiles have driven security researchers to caution using these tools to track children [114].

We found concerning privacy implications even in papers that didn't target children specifically. This raises concerns over how this kind of data collection could disproportionately impact other vulnerable populations. For example, papers [5, 48, 50, 51, 53–56, 76, 115, 146] collected data from sex work sites and provided this data to law enforcement and government entities. These datasets contain information on both human trafficking victims and voluntary sex workers; as a result, this data raises a number of privacy concerns beginning with the issue of misidentification. False positives generated by AI systems trained on these datasets could result in increased police and governmental intervention on groups historically affected by police and state violence. Further, these datasets tend to disproportionately contain data on marginalized identities – including BIPOC (Black, Indigenous People of Color) and LGBTQ+ people – who are already over-surveilled and subject to discriminatory practices resulting from increased surveillance [80, 126]. The use of satellite imagery as seen in papers [77] and [111] further sharpens our concerns. While tracking behaviors using satellite imagery have been used in a number of positive circumstances such as holding governments accountable and detecting cases of human rights abuses [17, 142], there are concerns that these systems will be used by nation-states to invasively monitor citizen's behaviors and could be used to systematically target marginalized groups [79].

In addition to data privacy, we as researchers must examine the ways publishing materials can potentially violate the right to privacy. In our corpus, most papers used mockups or synthetic data in examples to protect the privacy of victims. However, we found a few notable examples where researchers used real data within the text of the paper and/or used a public dataset with Personally identifiable information (PII) present. In all these cases, the PII concerned the perpetrators of sexual abuse or human trafficking. We caution against this practice and urge researchers to protect the privacy of all subjects in a dataset including perpetrators, even though the laws surrounding the privacy rights of those accused of crimes is in many cases unclear. This is especially important because even within the corpus how a person is being labeled a trafficker can be unclear (eg is a person labeled a trafficker upon conviction? After being accused? etc.). We further argue against releasing datasets without anonymizing this information and caution against using public datasets who engage in this practice as this has the potential to legitimize bad privacy practices (and in the case of one paper, potentially draw more attention to the social media accounts for suspected terrorists).

Finally, the issue of data access and sharing raises a number of questions concerning privacy. In our corpus, the datasets tended to fit into three categories:

(1) Public facing datasets that the authors did not create. For example, many papers use the IMDB-WIKI dataset [106] for training computer vision models for age estimation.
(2) self-collected datasets that an author created by scraping public webpages
(3) datasets provided to the authors by a third-party.

For public datasets, most papers used existing public-facing datasets and tended to have no data on human trafficking. This was particularly present in computer vision papers concerning age estimation who use human trafficking as a motivating reason to pursue this line of work but didn't exclusively examine human trafficking imagery. Researchers releasing public datasets about trafficking brings up the question of how their anonymization process guarantees privacy.

In cases where researchers collected their own datasets or were granted access to an existing private database, many of these datasets nominally collected no "identifiable data". Thus, it is unclear if laws such as GDPR[10] – which underlie most current human right approaches for protecting privacy rights – apply to many of the datasets described in this corpus. This brings up a number of questions concerning long-term data stewardship practices. Who is allowed to view the dataset including after the project is complete? What is the process for guaranteeing that only those individuals will have access to that data? Are there existing methods to proactively check for data-breaches? Long-term who retains control of these databases?

While answering these questions are typically is beyond the scope of most research papers, addressing these questions within publications represents an avenue for researchers to lead by example by demonstrating their principles for privacy protections and data stewardship. Additionally, by not addressing privacy concerns, it becomes difficult for reviewers and readers to evaluate the tool with respect to ethics.

6.2 Accountability

Accountability concerns the mechanisms that exist to monitor the impacts of AI systems over time as well as the processes to determine who is responsible when things go wrong. In practice, this principle has been applied to AI research through recommending new policy and regulations, requiring routine impact assessments, and establishing auditing requirements thorough an AI system's life-cycle.

However, conceptualizing accountability with respect to academic work can be challenging. In AI governance, Accountability is often discussed with respect to legal risk and responsibility. But this is not often applicable in academic settings because not all of the papers involve a fully deployed AI system. Thus, we instead focus our analysis on what mechanisms exist to address accountability aimed at the pre-deployment stage of AI development. This is most commonly through impact assessments which emphasize authors identifying and documenting potential harms and risks.

As all the papers in our corpus were peer-reviewed, we can assume that some level of this analysis was performed – ensuring a certain level of verifiability and replicability of the work. However, not all peer-review systems explicitly require reviewers to consider potential impacts and risks. Though this is becoming more common with more academic venues requiring authors to include analysis of harms and benefits in their work and to require reviewers to reflect on those points.

We found that documentation on the harms and benefits directly within the text of the papers was also inconsistent across the corpus. As mentioned in our findings section, very few papers explicitly mentioned issues of potential harm or identified risks. Additionally, few authors mention using impact assessment directly within the publications and even fewer authors address issues of long-term evaluations or mechanism for redress. There were some notable exceptions in our

[10] GDPR stands for the "General Data Protection Regulation" and this the European Union's data privacy and protection law

corpus however. [123] considered the specific needs of survivors and included specific design considerations to prevent data misuse, while [51] included an impact statement within the paper and emphasized design considerations for bias mitigation and redress.

6.3 Transparency and Explainability

Transparency and Explainability are concerned with the ability to understand and evaluate AI systems. Together, the relate to the idea that AI systems need to be designed so that oversight is possible and that technical concepts about an AI system should be translated into human-intelligible formats. Historically, this has been thought of as a binary (i.e. a system is either fully transparent or fully opaque), but more recent attention has been called to the idea that transparency – throughout the development life-cycle – exists on a continuum. Explainability is concerned with how technical concepts about an AI system can translated into human-intelligible formats.

Academic work can be transparent through open data and code, accessible documentation of the design process and results, and mechanisms to evaluate how the work will impact the general public. As discussed earlier in the privacy section, open data and code is not always a feasible option for research in this field. The need for transparency has to be balanced with the right to privacy and the need to protect investigative procedures. Though, some of the papers in the corpus have suggested dataset alternatives to ensure privacy while also promoting transparency. For example, some authors test their models using datasets that are related to the task but don't contain actual entries of human trafficking. For examples, papers [112] and [134] trained computer vision algorithms to match vein patterns in thighs and arms for forensic investigations of child pornography. Both papers use a dataset that collected images of these body parts from willing participants and not from sexual abuse imagery. Similarly, papers [9, 10, 28, 57] (which sought to improve age estimation of images for detecting cases of sex trafficking) online used public computer vision datasets of faces unrelated to trafficking to train their models. Similarly, work from related fields has suggested the use of "semi-synthetic" datasets as a workaround for this issue [143].

With respect to transparent documentation, all of the papers in the corpus clearly lay out their data collection, design process, and methods within the text. However, evaluating the transparency of this reporting is difficult when there are differing standards across publication venues for the reporting of results, limitations, etc. For example, only 20 papers (roughly one third) included a specific limitations section or the explicit discussion of limitations with another section such as the discussion section. Those that did discuss limitations tended to somewhat narrowly focus on dataset bias. Further, as mentioned in the accountability section, rarely did authors discuss the impact on stakeholders. The degree to which authors address potential harms and impact is somewhat determined by the norms of specific academic communities and page lengths.

Another issue affecting the transparency of this work is that academic papers are not the most accessible to all audiences. Almost all of the papers are hidden behind a paywall which limits who can easily access the information in the paper and evaluate the methods and limitations. Furthermore only 11 of the papers (roughly 15%) included an explicit statement about the availability of the data and/or code relevant to the study (including those that justified not making the data or code available) with only 9 of those papers including links to either the code or dataset. While some of the papers outside of these 11 used publicly available datasets, the authors did not mention this within the text on the paper.

6.4 Fairness and Non-Discrimination

Perhaps the most discussed topic within the corpus was the issue of bias. Most papers acknowledged the bias present in datasets, but none discussed bias in the model itself or in the focus of the applications. Discussions of bias are critical in this context as issues of discrimination in Human Trafficking applications can have profound life-or-death implications.

There are a number of areas where bias is potentially introduced into these systems, including bias present in the datasets, the labeling schemas, and the models trained as a result. There may also be bias present in what areas that computing researchers tended to focus on.

Within our corpus, we identified several incidents of bias with respect to gender, age, disability, and socioeconomic status that were present in the datasets that underlie these AI systems. In many cases these biases stems from a lack of sufficient training data. As papers from the corpus note [8, 28], many of the systems – especially those that employ facial recognition or age estimation techniques – lack sufficient training data on darker skin tones and thus are less accurate at identifying faces of BIPOC individuals. This issue of racial bias in computer vision applications is not unique to human trafficking efforts, but in this context these biases result in the development of real-world systems that are less accurate at identifying BIPOC victims and reinforces existing disparities in the criminal justice system. [8] further notes that many of the commercially available facial recognition algorithms used in identifying cases of child sex trafficking are less accurate on younger and female faces. This indicates that these systems – which are currently used today to specifically to identify cases of child sex trafficking – might not work as intended in that exact context. In practice, these biases translate to systems limiting which cases are investigated, which victims are identified, and which cases go to trial all while reinforcing existing racial and gender disparities in the criminal justice system.

It is also unclear how effective many of these computer vision models are on identifying faces with unique facial features such as those with Down's Syndrome, facial scarring, or prominent birthmarks. Given that none of the training datasets explicitly mentioned collecting images of those with disabilities or facial differences, it is likely that the accuracy of these models are limited in this area. Individuals with disabilities are at increased risk of exploitation [92]. Again these systems have the effect of steering investigations towards only certain groups.

Related to this issue of dataset bias, is the issue of gathering valid and representative "ground truth labels" as to a person's status as a victim of human trafficking or relating to a person's likelihood of being trafficked in the future. In this space, the most common application of AI is for automatic identification of human trafficking victims and/or the prediction of the likelihood of a person being trafficked in the future. Ground truth data labels are directly fed into most systems and used to measure the accuracy of models developed. Thus, establishing the validity of these labels is a critical process to research in this space and directly impacts the quality and effectiveness of the resulting systems. We identified a number of challenges in gathering ground truth labels revealed by our corpus. First, there is a general lack of consensus on a clear definition of human trafficking. As a result, some authors conflate instances of voluntary work with human trafficking or conflate certain behaviors as being high risk for trafficking. For example, one paper collected data from an online fetish website and conflated interest in consensual taboo sexual preferences for involuntary human trafficking. Compounding this is the general "fuzziness" with distinguishing between exploitation and trafficking. As many researchers note, human trafficking exists on a spectrum and drawing a firm line is a complicated task [74]. Even law enforcement are reluctant to make firm judgements on what is or isn't human trafficking - saying that it takes multiple in-person interviews to determine that [35]. Thus, this process of labeling ground truth injects bias into the dataset as these labels will reflect the cultural values of the researchers and experts assisting with labeling. Additionally, the models themselves may learn racist human trafficking indicators was the case with [5] which found that their machine learning model learned that the phrase "Asian" indicates sex trafficking.

There is also a general lack of established ground truth data sets - especially with respect to distinguishing between consensual sex work and human trafficking. There are no standardized techniques to label activity as human trafficking or as high risk of trafficking. Thus, researchers rely on proxy indicators of trafficking such as suspicious key words that are suggested to be indicative of trafficking. Most commonly, researchers have used law enforcement or NGO-generated key-words to label advertisements as likely involving human trafficking as seen in [5, 6, 51, 53–56, 76, 99]. However, experts often disagree as to how accurate these key-words are at predicting human trafficking cases [62, 74, 91, 125]. As one researcher put it “no researcher or investigator can ascertain with 100% confidence that a particular online advertisement is a positive case of sex trafficking, just as one cannot be completely certain whether an advertisement is a negative case" [74]. To date no research has empirically evaluated the reliability of these keywords.

Finally, as noted earlier in our results section, the overwhelming majority of papers focused on sex trafficking - with a nearly even split of adult and child cases – and focused heavily on building systems for investigative purposes. There is a serious gap in work addressing labor trafficking and work that addresses the needs of survivors and service providers. The principles of fairness and non-discrimination extend beyond algorithmic bias and also include the principles of inclusiveness in design and in impact. Only two papers took into account the needs of survivors [42, 123] and none explicitly included survivors as authors. Only a few papers mentioned working directly with their intended users as part of the design process.

6.5 Human Control of Technology

The principle of Human Control of Technology deals with the ethical tensions that arise as a result of shifts in control away from people to AI systems. Both the individuals who use an AI system and those who are impacted by said system should be able to review decisions and remedy any objectionable results. In practice, human control of technology takes on a variety of forms – such as designs that integrate ex post expert reviews, models that incorporate direct human input, and even designs that include models that predict when humans should intervene.

In our corpus, we found a spectrum of how authors included human input/reviews – from fully autonomous designs where expert input was limited to labeling and evaluation tasks, to designs that fully integrated direct human-input. Where a paper fell on this spectrum tended to depend heavily on what computing sub-discipline the authors drew their methods from. Papers using methods from data science and machine learning tended only include human input in dataset labeling and model evaluation tasks, whereas papers who used methods from HCI tended to include more involved human-in-the-loop elements in their designs. However, the overwhelming majority of papers in our corpus limited human input to only dataset labeling and rarely included expert evaluation of the results after the fact. Papers [5, 38, 52, 53, 55, 76, 77, 112, 122] all used experts to label their training data and used these labels to evaluate the accuracy of their model’s decisions; but none of these papers included expert evaluation of these decisions afterwards. This gap highlights opportunities for future research to include more avenues for human interaction.

This also brings up the concern that incorporating AI systems is inherently changing human behavior. Rather than human users acting as safeguards, these systems may influence their users to instead change behaviors [144]. Further research is needed to understand how in-situ uses of AI are affecting excising decision making process, especially in the context of criminal justice systems and policy oversight.

6.6 Professional Responsibility

This principle emphasizes the personal responsibility researchers have when designing any AI system and argues that researchers should ensure that appropriate stakeholders are included and that long-term impacts are accounted for. In our analysis, we examined this principle with respect to themes mentioned in the framework: 1) responsible design, 2) consideration for long-term effects, and 3) multi-sector collaboration.

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 papers represent snap-shots. However, a somewhat troubling trend is how few papers directly mention long-term plans or the evaluation of future harms. Related to the discussion of long-term effects, is the also question of long-term data stewardship. What happens to these datasets after a research project is complete? What should the long-term data plan be for projects using human trafficking data?

Finally, multi-sector collaboration is a necessity with this research. However beyond labeling and data access, we need to ask what it means to partner with external groups on research projects. These partnerships can be a mechanism to include voices typically excluded in academic settings. But note that these groups are not necessarily bound by the same standards as academics; there are little regulations with how private organization operate, collect data, or use results from research. Additionally, this research could also be unintentionally lending credibility to harmful industries and practices. For example, [110] partnered with an organization comprised of "psychic detectives", a well-known predatory industry.

6.7 Promotion of Human Values

This principle is concerned with the idea that AI should be developed and used to promote "human flourishing" and leveraged to benefit society. All of the papers in our corpus intended to do this. The goal of these papers is to use computing to combat human trafficking and support existing anti-trafficking efforts. They also represent avenues to inform the broader public of the issue of human trafficking.

However, given recent controversies surrounding law enforcement use of AI and its ability to cause harm, it is worth examining how these tools might be used in ways that could be detrimental to society. Facial recognition is an interesting case study to examine in this context - especially as many of the papers are either developing systems for facial recognition or use facial recognition as a component within a tool used by law enforcement. We recognize that facial recognition is an important tool for law enforcement; It has, for example, been used to search for missing persons suspected to be trafficked, identify perpetrators in child abuse imagery, and forensically link evidence over time. There have been several notable cases where the use of commercially available facial recognition APIs have lead to the arrest of a trafficker.

However, these same APIs have also been used by law enforcement and governments in other contexts with little oversight. For example, during the US Black Lives Matter Protests in 2015 and 2020, facial recognition was used by law enforcement to identify and arrest protesters [104]. In China, facial recognition has long been a component of the systematic processes used to surveil and oppress ethnic minorities [30]. Additionally, because of bias present in the models and data, facial recognition has resulted in a number of wrongful arrests caused solely by algorithmic miss-identification [7]. Together these concerns have lead many civil rights organizations to call for the ban of facial recognition used for law enforcement purposes until better safeguards can be put in place [104].

This brings up the concern that the development of AI for human trafficking purposes can cause unintended harm especially towards marginalized communities. More research is needed to understand the impact of AI in criminal justice and governmental settings. This research should further include mechanisms for community input and control with particular attention given to including marginalized communities.

7 DISCUSSION

As research in this area continues to grow, we as a research community should consider how this work will affect the world around us. We believe this paper marks a first-step towards examining the ethical tensions present in this line of work and further highlights areas for future research. In this section, we briefly summarize some of the challenges we identified and propose the following calls to action and future research directions.

7.1 Broader Use of Participatory Design

Responding to the problem of human trafficking necessarily requires interdisciplinary and multi-sector collaboration. The research community has already responded to this need through partnerships with NGOs, law enforcement, the civil sector as well as through cross-discipline academic research. However as discussed in our analysis, there is a need for researchers to include a broader range of stakeholders in the research process – with special care taken to include those most impacted by the research. AI research on human trafficking often impacts marginalized groups including sex workers, migrant workers, queer identities, and people of color. In addition, data collection has the potential to further harm and stigmatize survivors of human trafficking.

Despite how this work may impact these groups, we found that few papers directly included survivors, sex workers, or migrant workers in the research process. There is a strong need for future work to directly incorporate a broader range of stakeholders within the research process. As the products of AI research can further marginalize already vulnerable groups, future AI development for human trafficking should ensure that survivors are involved in the development and research process. Participation from these groups can take many forms: survivors could be involved to vet ground-truth data used to train these systems and evaluate research outcomes, through processes to evaluate and inform research questions and applications, and through direct inclusion as researchers themselves. Future work could also benefit from action research performed in partnership with survivor groups. Prior work (found both in our corpus[42, 123] and elsewhere[129]) has shown that partnerships with survivor groups helps center survivor voices in the research.

However, how participation is included in practice will have profound effects on future AI development. Participatory research is not without its flaws [67, 89] and we encourage researchers to examine the power structures embedded in participation [129]. Particularly when including survivors in the research process, we must be cognizant of the fact that participation must provide benefits to all participants. Some survivors who have shared their experiences have felt exploited by the researcher and anti-trafficking community [32, 129]. Researchers need to ensure that survivors are not further exploited by the research process and that survivors are recognized and compensated for their expertise and knowledge. The goal with engaging with practices like action research and participatory research should be emancipation and democratization [47]. To this end, we argue that survivors and other marginalized groups should be directly and equally included in the research process – not as subjects or users, but instead as project leaders, researchers, and PhD students. Towards this goal of direct inclusion, future work should also examine what existing barriers are in place that have historically prevented direct and equal participation from these groups. This represents an opportunity for researchers to examine the ways in which technology can further participation and cross-disciplinary research with under-served populations.

7.2 Broadening Research to Address Gaps and Engage with Other Forms of Trafficking

As discussed earlier in the "Fairness and Non-Discrimination" section, there are a number of areas that have seen little attention from the research community. Thus far, the research community has disproportionately focused on sex trafficking and in particular child sex trafficking. As a result, there is a distinctive lack of research addressing labor trafficking – despite estimates pointing towards labor trafficking being the most common form of trafficking worldwide [58]. Further within research specifically focusing on sex trafficking, there is a lack of research addressing sex trafficking experiences beyond forced prostitution and pornography. None of the papers we have reviewed have addressed the issue of forced marriage and only one of the papers [31] addressed cybersex trafficking. Additionally, the existing work on sex trafficking tends to focus heavily on women, girls, and trans-feminine victims. Notably, many of the data sources used by researchers exclude men and trans-masculine individuals; thus to date, little research has focused specifically on addressing sex trafficking of men and non-binary individuals. Finally, the existing work has more broadly focused on assisting "prosecution"-aligned efforts and in particular on victim identification rather than addressing the other 3Ps: Protection, Prevention, Partnership.

Together, these gaps highlight the need for future work to examine the larger human trafficking ecosystem. Prior work has noted that labor trafficking is increasingly intersecting with technology which represents avenues for researchers to understand the labor trafficking ecosystem using similar methods as the work done to understand the sex trafficking ecosystem. These gaps also highlight the need for further research on the ways in which technology is changing sex trafficking and mechanisms to prevent internet-facilitated exploitation. We urge researchers to use their platform and audience to draw attention to other forms of trafficking.

In addition to broadening the focus of AI research, we also advocate for further research focusing on the impacts of deploying this research. Particularly as much of the work intersects with the criminal justice system, we urge researchers to examine how these tools are impacting decision-making processes. We call for more research focused on examining the use of AI in-situ perhaps using similar approaches as those used in understanding AI impacts in healthcare, etc [44].

7.3 Development of Best Practices for Harm Prevention

Research directed towards human trafficking applications present unique avenues for potential harm. There is the potential to cause harm through the data collection process, through the design choices made by researchers, through misuse of existing tools, and even through harm to the researchers caused from interacting with disturbing materials. To this end, we urge the research community to form a series of best practices targeting harm prevention with a particular focus on data privacy and security practices, long-term data stewardship practices, and transparent design documentation. Careful attention must be paid towards ensuring that research efforts do not cause more harm than good and that the work does not infringe upon the rights of individuals impacted by the research. When developing these best practices, the research community should consult with those most impacted by this research – including survivors, migrant workers, and sex workers – to ensure that these best practices reflect their needs.

In addition, these best practices can draw from guidelines developed for computing research that handles similarly sensitive topics. Borrowing from guidelines developed for computing research that handles sensitive data, future work should include mechanisms in their designs for proactive monitoring, use data sharing agreements that limit access to sensitive data, and include policies to delete datasets after a project is complete [16, 27]. With respect to data privacy, future work should consider using mock datasets alongside other datasets to improve the transparency and accountability of the work while also preserving the privacy of survivors. Finally throughout the design process and within the text of the publication, researchers should document potential harms, benefits, and most importantly strategies for mitigating harm.

7.4 Broader Inclusion of Ethics Disclosures in Research and Transparent Discussions of Limitations

Our analysis highlighted the need for a standardized practice of including ethics discussions in academic work and for researchers to include transparent and accessible documentation of terms of use and limitations of their work. The practice of including ethics discussion – particularly an ethics discussion focusing on the potential harms and benefits – would help ensure that researchers consider the impacts of their work and would further help the broader community asses the impact of these AI systems. However, there is a real asymmetry in power and knowledge between those who create these AI systems and those who use and are impacted by said systems. Researchers must be cognizant of this asymmetry in designing and writing their limitations and careful attention must be paid to ensuring that the broader public fully understands the impact of their work.

Researchers should also include similar discussion of limitations and biases for their datasets. Especially for datasets that are shared beyond those who collected it, datasets need to include documentation detailing the data's provenance, limitations for usage, and any known biases. We encourage researchers to consider using tools such as the one provided by the Data Nutrition Project[11] to detail these data limitations in an accessible format [29].

Finally, researchers should consider ways to include the discussion of limitations within the design of the tool itself. Tool interfaces could, for example, include disclaimers written in clear language detailing the limitations of the tool. Additionally drawing from fields like information visualization, future research could look at how can we can design interfaces so that users can clearly understand the limitations of an AI system and the biases of the datasets used [137, 138].

7.5 Limitations of AI for Human Trafficking

Human trafficking is inherently linked to inequality and marginalization [92, 136]. Human trafficking is a deeply complex problem that is difficult to research and requires navigating intersecting complex social, economic, cultural, and political structures that facilitate exploitation. At the root of it all, those most vulnerable to exploitation are marginalized groups who have unequal economic opportunities. Thus, dismantling systems of oppression and implementing necessary social programs (like increasing the number of long-term shelters and support for equal education and employment opportunities) will have profound effects towards ending human trafficking. To this end, we must be cognizant of where and when AI should and should not be used. Technology and AI systems should not replace or be used instead of much needed social programs, training, and education. We must be careful as computing researchers not to divert resources from social programs towards technology development.

Further, AI systems should not be marketed as “bias-free" or “objective" tools that overcome the limitations of human decision making. Especially in the context of criminal justice applications, AI systems have been shown time and time again to both reproduce and magnify existing inequalities [21]. Researchers must pay attention to the language they use when describing their tool's intended use cases to watch for descriptions that imply this. As discussed earlier, the limitations of any system must be understood by its users.

[11] https://datanutrition.org/

8 CONCLUSION

With the availability of new datasets and the increased awareness, applying AI to combat human trafficking is an emerging area for new research. In this paper, we analyzed a corpus on this topic using human rights as a lens to highlight ethical tensions that arise from this work. We further propose five calls to action that highlight avenues for ethically-driven future work. We hope this work provides inspiration for thoughtful future research aimed at tackling human trafficking.

REFERENCES

[1] [n.d.]. *Regional Trial Court of Misamis Oriental, the Philippines, 10th Judicial Region, Branch 41, CRIM Case NO. 2009-337.*

[2] 2000. United States of America: Victims of Trafficking and Violence Protection Act of 2000. Public Law 106-386 [H.R. 3244].

[3] Evgeni Aizenberg and Jeroen van den Hoven. 2020. Designing for human rights in AI. *Big Data & Society* 7, 2 (2020), 2053951720949566.

[4] Hamdan Z. Alshammari and Khaled S. Alghathbar. 2017. CLogVis: Crime Data Analysis and Visualization Tool. In *Proceedings of the Second International Conference on Advanced Wireless Information, Data, and Communication Technologies* (Paris, France) *(AWICT 2017)*. Association for Computing Machinery, New York, NY, USA, Article 2, 7 pages. https://doi.org/10.1145/3231830.3231832

[5] Hamidreza Alvari, Paulo Shakarian, and JE Kelly Snyder. 2016. A non-parametric learning approach to identify online human trafficking. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*. IEEE, 133–138.

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

[7] Amnesty International. [n.d.]. Ban dangerous facial recognition technology that amplifies racist policing. https://www.amnesty.org/en/latest/news/2021/01/ban-dangerous-facial-recognition-technology-that-amplifies-racist-policing/#:~:text=While%20other%20US%20cities%2C%20including,the%20Black%20Lives%20Matters%20protests.

[8] Felix Anda, Brett A Becker, David Lillis, Nhien-An Le-Khac, and Mark Scanlon. 2020. Assessing the Influencing Factors on the Accuracy of Underage Facial Age Estimation. In *2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security)*. IEEE, 1–8.

[9] Felix Anda, David Lillis, Aikaterini Kanta, Brett A. Becker, Elias Bou-Harb, Nhien-An Le-Khac, and Mark Scanlon. 2019. Improving Borderline Adulthood Facial Age Estimation through Ensemble Learning. In *Proceedings of the 14th International Conference on Availability, Reliability and Security* (Canterbury, CA, United Kingdom) *(ARES '19)*. Association for Computing Machinery, New York, NY, USA, Article 57, 8 pages. https://doi.org/10.1145/3339252.3341491

[10] Felix Anda, David Lillis, Nhien-An Le-Khac, and Mark Scanlon. 2018. Evaluating automated facial age estimation techniques for digital forensics. In *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 129–139.

[11] Lindsey Andersen et al. 2018. Human Rights in the age of Artificial Intelligence. *Access Now* (2018), 29.

[12] Simon Andrews, Ben Brewster, and Tony Day. 2018. Organised crime and social media: a system for detecting, corroborating and visualising weak signals of organised crime online. *Security Informatics* 7, 1 (2018), 1–21.

[13] UN General Assembly et al. 1948. Universal declaration of human rights. *UN General Assembly* 302, 2 (1948), 14–25.

[14] Paul Baker and Amanda Potts. 2013. 'Why do white people have thin lips?'Google and the perpetuation of stereotypes via auto-complete search forms. *Critical discourse studies* 10, 2 (2013), 187–204.

[15] G. Josemin Bala and Steven Lawrence Fernandes. 2016. Developing Novel Skin Detection on ODROID XU4 Heterogeneous Multi-Processing Device. In *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies* (Udaipur, India) *(ICTCS '16)*. Association for Computing Machinery, New York, NY, USA, Article 86, 4 pages. https://doi.org/10.1145/2905055.2905297

[16] Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. 94–102.

[17] Chris Beyrer and Adeeba Kamarulzaman. 2017. Ethnic cleansing in Myanmar: the Rohingya crisis and human rights. *The Lancet* 390, 10102 (2017), 1570–1573.

[18] Andrew Booth, Anthea Sutton, and Diana Papaioannou. 2016. Systematic approaches to a successful literature review. (2016).

[19] Vanessa Bouche et al. 2015. A report on the use of technology to recruit, groom and sell domestic minor sex trafficking victims. (2015).

[20] Danah Boyd and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society* 15, 5 (2012), 662–679.

[21] Sarah Brayne. 2020. *Predict and surveil: Data, discretion, and the future of policing.* Oxford University Press, USA.

[22] Jeffrey Brelsford and Saurabh Parakh. 2018. A systems modeling approach to analyzing human trafficking. In *2018 Winter Simulation Conference (WSC)*. IEEE, 12–21.

[23] Morten Broberg and Hans-Otto Sano. 2018. Strengths and weaknesses in a human rights-based approach to international development–an analysis of a rights-based approach to development assistance based on practical experiences. *The International Journal of Human Rights* 22, 5 (2018), 664–680.

[24] Axel Bruns, Katrin Weller, Michael Zimmer, and Nicholas John Proferes. 2014. A topology of Twitter research: disciplines, methods, and ethics. *Aslib Journal of Information Management* (2014).

[25] Danilo Burbano and Myriam Hernández-Alvarez. 2018. Illicit, hidden advertisements on Twitter. In *2018 International Conference on eDemocracy & eGovernment (ICEDEG)*. IEEE, 317–321.

[26] Joshua Carback. 2018. Cybersex Trafficking: Toward a More Effective Prosecutorial Response. *Criminal Law Bulletin* 54, 1 (2018).

[27] Stevie Chancellor, Michael L Birnbaum, Eric D Caine, Vincent MB Silenzio, and Munmun De Choudhury. 2019. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the conference on fairness, accountability, and transparency*. 79–88.

[28] Praveen Kumar Chandaliya and Neeta Nain. 2019. Conditional Perceptual Adversarial Variational Autoencoder for Age Progression and Regression on Child Face. In *2019 International Conference on Biometrics (ICB)*. IEEE, 1–8.

[29] Kasia S Chmielinski, Sarah Newman, Matt Taylor, Josh Joseph, Kemi Thomas, Jessica Yurkofsky, and YC Qiu. 2020. The Dataset Nutrition Label (2nd Gen): Leveraging Context to Mitigate Harms in Artificial Intelligence. In *Proceedings of the NeurIPS 2020 Workshop on Dataset Curation and Security, Online*, Vol. 11.

[30] Niraj Chokshi. 2019. Facial Recognition's Many Controversies, From Stadium Surveillance to Racist Software'. *New York Times* 15 (2019).

[31] Wingyan Chung, Elizabeth Mustaine, and Daniel Zeng. 2017. Criminal intelligence surveillance and monitoring on social media: Cases of cyber-trafficking. In *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 191–193.

[32] Karen Countryman-Roswurm and Bailey Patton Brackin. 2017. Awareness without re-exploitation: Empowering approaches to sharing the message about human trafficking. *Journal of Human Trafficking* 3, 4 (2017), 327–334.

[33] Tessa Couture. 2016. *More than Drinks for Sale: Exposing Sex Trafficking in Cantinas & Bars in the U.S.* Technical Report. Polaris.

[34] M. da Silva Santos, M. Ladeira, G. C. G. Van Erven, and G. Luiz da Silva. 2019. Machine Learning Models to Identify the Risk of Modern Slavery in Brazilian Cities. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. 740–746.

[35] Julia Deeb-Swihart, Alex Endert, and Amy Bruckman. 2019. Understanding law enforcement strategies and needs for combating human trafficking. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[36] Matthew Edwards, Awais Rashid, and Paul Rayson. 2015. A systematic survey of online data mining technology intended for law enforcement. *ACM Computing Surveys (CSUR)* 48, 1 (2015), 1–54.

[37] Paul Elzinga, Karl Erich Wolff, and Jonas Poelmans. 2012. Analyzing chat conversations of pedophiles with temporal relational semantic systems. In *2012 European Intelligence and Security Informatics Conference*. IEEE, 242–249.

[38] Muhammad Ali Fauzi and Patrick Bours. 2020. Ensemble Method for Sexual Predators Identification in Online Chats. In *2020 8th International Workshop on Biometrics and Forensics (IWBF)*. IEEE, 1–6.

[39] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication* 2020-1 (2020).

[40] Sakiko Fukuda-Parr and Elizabeth Gibbons. 2021. Emerging Consensus on 'Ethical AI': Human Rights Critique of Stakeholder Guidelines. *Global Policy* 12 (2021), 32–44.

[41] Aakash Gautam, Chandani Shrestha, Andrew Kulak, Steve Harrison, and Deborah Tatar. 2018. Participatory tensions in working with a vulnerable population. In *Proceedings of the 15th Participatory Design Conference: Short Papers, Situated Actions, Workshops and Tutorial-Volume 2*. 1–5.

[42] Aakash Gautam, Deborah Tatar, and Steve Harrison. 2020. Crafting, Communality, and Computing: Building on Existing Strengths To Support a Vulnerable Population. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376647

[43] N. A. Giacobe, J. B. Altmire, A. E. Forster, A. C. Jackson, E. W. Raibick, J. A. Reep, R. Y. Tsang, and P. K. Forster. 2016. Characterizing sex trafficking in Pennsylvania for law enforcement. In *2016 IEEE Symposium on Technologies for Homeland Security (HST)*. 1–5.

[44] Tarleton Gillespie and Nick Seaver. 2016. Critical algorithm studies: A reading list. *Social Media Collective* (2016).

[45] Mitchell Goist, Ted Hsuan Yun Chen, and Christopher Boylan. 2019. Reconstructing and analyzing the transnational human trafficking network. In *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 493–500.

[46] Sergio L Granizo, Ángel Leonardo Valdivieso Caraguay, Lorena Isabel Barona López, and Myriam Hernández-Álvarez. 2020. Detection of Possible Illicit Messages Using Natural Language Processing and Computer Vision on Twitter and Linked Websites. *IEEE Access* 8 (2020), 44534–44546.

[47] Gillian R Hayes. 2020. Inclusive and engaged HCI. *Interactions* 27, 2 (2020), 26–31.

[48] M. Hernández-Álvarez. 2019. Detection of Possible Human Trafficking in Twitter. In *2019 International Conference on Information Systems and Software Technologies (ICI2ST)*. 187–191.

[49] AI HLEG. 2019. High-level Expert Group on Artificial Intelligence: Ethics guidelines for trustworthy AI. *European Commission, 09.04* (2019).

[50] Marisa Hultgren, Murray E Jennex, John Persano, and Cezar Ornatowski. 2016. Using knowledge management to assist in identifying human sex trafficking. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*. IEEE, 4344–4353.

[51] Kyle Hundman, Thamme Gowda, Mayank Kejriwal, and Benedikt Boecking. 2018. Always Lurking: Understanding and Mitigating Bias in Online Human Trafficking Detection. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New Orleans, LA, USA) *(AIES '18)*. Association for Computing Machinery, New York, NY, USA, 137–143. https://doi.org/10.1145/3278721.3278782

[52] Ryan Hurley, Swagatika Prusty, Hamed Soroush, Robert J. Walls, Jeannie Albrecht, Emmanuel Cecchet, Brian Neil Levine, Marc Liberatore, Brian Lynn, and Janis Wolak. 2013. Measurement and Analysis of Child Pornography Trafficking on P2P Networks. In *Proceedings of the 22nd International Conference on World Wide Web* (Rio de Janeiro, Brazil) *(WWW '13)*. Association for Computing Machinery, New York, NY, USA, 631–642. https://doi.org/10.1145/2488388.2488444

[53] M. Ibanez and R. Gazan. 2016. Detecting sex trafficking circuits in the U.S. through analysis of online escort advertisements. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 892–895.

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

[55] M. Ibanez and D. D. Suthers. 2014. Detection of Domestic Human Trafficking Indicators and Movement Trends Using Content Available on Open Internet Sources. In *2014 47th Hawaii International Conference on System Sciences*. 1556–1565.

[56] M. Ibanez and D. D. Suthers. 2016. Detecting covert sex trafficking networks in virtual markets. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 876–879.

[57] Amin Ibrahim and Miguel Vargas Martin. 2009. Addressing privacy constraints for efficient monitoring of network traffic for illicit images. In *2009 IEEE Toronto International Conference Science and Technology for Humanity (TIC-STH)*. IEEE, 302–308.

[58] International Labour Office. 2017. *Global Estimates of Modern Slavery: Forced Labour and Forced Marriage*. http://www.ilo.org/global/publications/books/WCMS_575479/lang--en/index.htm

[59] Lucas Introna and David Wood. 2004. Picturing algorithmic surveillance: The politics of facial recognition systems. *Surveillance & Society* 2, 2/3 (2004), 177–198.

[60] Rikke Frank Jørgensen. 2019. *Human rights in the age of platforms*. The MIT Press.

[61] Rahul Kapoor, Mayank Kejriwal, and Pedro Szekely. 2017. Using Contexts and Constraints for Improved Geotagging of Human Trafficking Webpages. In *Proceedings of the Fourth International ACM Workshop on Managing and Mining Enriched Geo-Spatial Data* (Chicago, Illinois) *(GeoRich '17)*. Association for Computing Machinery, New York, NY, USA, Article 3, 6 pages. https://doi.org/10.1145/3080546.3080547

[62] Mayank Kejriwal, Jiayuan Ding, Runqi Shao, Anoop Kumar, and Pedro Szekely. 2017. Flagit: A system for minimally supervised human trafficking indicator mining. *arXiv preprint arXiv:1712.03086* (2017).

[63] Mayank Kejriwal and Yao Gu. 2020. Network-theoretic modeling of complex activity using UK online sex advertisements. *Applied Network Science* 5, 1 (2020), 1–23.

[64] Mayank Kejriwal and Rahul Kapoor. 2019. Network-theoretic information extraction quality assessment in the human trafficking domain. *Applied Network Science* 4, 1 (2019), 1–26.

[65] Mayank Kejriwal and Pedro Szekely. 2017. Information Extraction in Illicit Web Domains. In *Proceedings of the 26th International Conference on World Wide Web* (Perth, Australia) *(WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 997–1006. https://doi.org/10.1145/3038912.3052642

[66] Mayank Kejriwal and Pedro Szekely. 2018. Technology-Assisted Investigative Search: A Case Study from an Illicit Domain. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/

3170427.3174364

[67] Finn Kensing and Jeanette Blomberg. 1998. Participatory design: Issues and concerns. *Computer supported cooperative work (CSCW)* 7, 3 (1998), 167–185.

[68] Reuben Kirkham. 2020. Using European Human Rights Jurisprudence for Incorporating Values into Design. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 115–128.

[69] Barbara Kitchenham and Stuart Charters. 2007. Guidelines for performing systematic literature reviews in software engineering. (2007).

[70] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.

[71] Renata A Konrad, Andrew C Trapp, Timothy M Palmbach, and Jeffrey S Blom. 2017. Overcoming human trafficking via operations research and analytics: Opportunities for methods, models, and applications. *European Journal of Operational Research* 259, 2 (2017), 733–745.

[72] Panos Kostakos, Lucie Špráchalová, Abhinay Pandya, Mohamed Aboeleinen, and Mourad Oussalah. 2018. Covert online ethnography and machine learning for detecting individuals at risk of being drawn into online sex work. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 1096–1099.

[73] PM Krafft, Meg Young, Michael Katell, Karen Huang, and Ghislain Bugingo. 2020. Defining AI in policy versus practice. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 72–78.

[74] Mark Latonero. 2011. Human trafficking online: The role of social networking sites and online classifieds. *Available at SSRN 2045851* (2011).

[75] Mark Latonero. 2018. Governing artificial intelligence: Upholding human rights & dignity. *Data & Society* (2018), 1–37.

[76] Lin Li, Olga Simek, Angela Lai, Matthew Daggett, Charlie K Dagli, and Cara Jones. 2018. Detection and characterization of human trafficking networks using unsupervised scalable text template matching. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 3111–3120.

[77] Xiaodong Li, Giles M Foody, Doreen S Boyd, and Feng Ling. 2019. Aging brick kilns in the asian brick belt using a long time series of Landsat sensor data to inform the study of modern day slavery. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 130–133.

[78] Marc Liberatore, Brian Neil Levine, and Clay Shields. 2010. Strengthening Forensic Investigations of Child Pornography on P2P Networks. In *Proceedings of the 6th International COnference* (Philadelphia, Pennsylvania) *(Co-NEXT '10)*. Association for Computing Machinery, New York, NY, USA, Article 19, 12 pages. https://doi.org/10.1145/1921168.1921193

[79] Amos Lichtman and Mohit Nair. 2015. Humanitarian uses of drones and satellite imagery analysis: the promises and perils. *AMA journal of ethics* 17, 10 (2015), 931–937.

[80] David Lyon. 2001. *Surveillance society: Monitoring everyday life*. McGraw-Hill Education (UK).

[81] Marinus Analytics. 2019. TRAFFIC JAM for FACIAL RECOGNITION: Sex Trafficking Victim Found Online from 2 Year Old Photo.

[82] Estelle Massé. 2020. *Recommendations on privacy and data protection in the fight against COVID-19*. Technical Report. Access Now.

[83] Lorna McGregor, Daragh Murray, and Vivian Ng. 2019. International human rights law as a framework for algorithmic accountability. *International & Comparative Law Quarterly* 68, 2 (2019), 309–343.

[84] Jacob Metcalf and Kate Crawford. 2016. Where are human subjects in big data research? The emerging ethics divide. *Big Data & Society* 3, 1 (2016), 2053951716650211.

[85] Dimitrios Michalopoulos and Ioannis Mavridis. 2010. Towards risk based prevention of grooming attacks. In *2010 International Conference on Security and Cryptography (SECRYPT)*. IEEE, 1–4.

[86] Dimitrios Michalopoulos and Ioannis Mavridis. 2011. Utilizing document classification for grooming attack recognition. In *2011 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 864–869.

[87] Dimitrios Michalopoulos, Eustathios Papadopoulos, and Ioannis Mavridis. 2012. Artemis: protection from sexual exploitation attacks via SMS. In *2012 16th Panhellenic Conference on Informatics*. IEEE, 19–24.

[88] Josephina Mikka-Muntuumo, Anicia Peters, and Hussin Jazri. 2018. CyberBullet - Share Your Story: An Interactive Game for Stimulating Awareness on the Harm and Negative Effects of the Internet. In *Proceedings of the Second African Conference for Human Computer Interaction: Thriving Communities* (Windhoek, Namibia) *(AfriCHI '18)*. Association for Computing Machinery, New York, NY, USA, Article 54, 4 pages. https://doi.org/10.1145/3283458.3283482

[89] Meredith Minkler. 2004. Ethical challenges for the "outside" researcher in community-based participatory research. *Health Education & Behavior* 31, 6 (2004), 684–697.

[90] Kimberly J Mitchell and Dana Boyd. 2014. Understanding the role of technology in the commercial sexual exploitation of children: the perspective of law enforcement. (2014).

[91] Jessica D Moorman and Kristen Harrison. 2016. Gender, race, and risk: Intersectional risk management in the sale of sex online. *The Journal of Sex Research* 53, 7 (2016), 816–824.

[92] Andrea J Nichols and Andrea Nichols. 2017. Sex Trafficking in the United States. In *Sex Trafficking in the United States*. Columbia University Press.

[93] Helen Nissenbaum. 2001. How computer systems embody values. *Computer* 34, 3 (2001), 120–119.

[94] US Department of State. 2010. Four "Ps": Prevention, protection, prosecution, partnerships.

[95] Chitu Okoli. 2015. A guide to conducting a standalone systematic literature review. *Communications of the Association for Information Systems* 37, 1 (2015), 43.

[96] Kien Pham, Aécio Santos, and Juliana Freire. 2018. Learning to Discover Domain-Specific Web Content. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (Marina Del Rey, CA, USA) *(WSDM '18)*. Association for Computing Machinery, New York, NY, USA, 432–440. https://doi.org/10.1145/3159652.3159724

[97] Jason Pielemeier. 2019. AI & Global Governance: The Advantages of Applying the International Human Rights Framework to Artificial Intelligence. *Centre for Policy Research at United Nations University. Retrieved* 30 (2019).

[98] Jonas Poelmans, Paul Elzinga, Stijn Viaene, and Guido Dedene. 2010. Concept discovery innovations in law enforcement: A perspective. In *2010 International Conference on Intelligent Networking and Collaborative Systems*. IEEE, 473–478.

[99] Rebecca S. Portnoff, Danny Yuxing Huang, Periwinkle Doerfler, Sadia Afroz, and Damon McCoy. 2017. Backpage and Bitcoin: Uncovering Human Traffickers. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS, Canada) *(KDD '17)*. Association for Computing Machinery, New York, NY, USA, 1595–1604. https://doi.org/10.1145/3097983.3098082

[100] Reihaneh Rabbany, David Bayani, and Artur Dubrawski. 2018. Active Search of Connections for Case Building and Combating Human Trafficking. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 2120–2129. https://doi.org/10.1145/3219819.3220103

[101] Vinoth Rengaraj and Kamal Bijlani. 2016. A study and implementation of smart ID card with M-learning and child security. In *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*. IEEE, 305–311.

[102] Claire M Ribando. 2007. Trafficking in Persons: US policy and issues for Congress. LIBRARY OF CONGRESS WASHINGTON DC CONGRESSIONAL RESEARCH SERVICE.

[103] Tatiana R Ringenberg, Kanishka Misra, and Julia Taylor Rayz. 2019. Not so cute but fuzzy: Estimating risk of sexual predation in online conversations. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 2946–2951.

[104] Katelyn Ringrose. 2019. Law Enforcement's Pairing of Facial Recognition Technology with Body-Worn Cameras Escalates Privacy Concerns. *Va. L. Rev. Online* 105 (2019), 57.

[105] S. Roshan, S. V. Kumar, and M. Kumar. 2017. Project spear: Reporting human trafficking using crowdsourcing. In *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)*. 295–299.

[106] Rasmus Rothe, Radu Timofte, and Luc Van Gool. 2018. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision* 126, 2-4 (2018), 144–157.

[107] Jennifer Salerno, Bartha M Knoppers, Lisa M Lee, Wayway M Hlaing, and Kenneth W Goodman. 2017. Ethics, big data and computing in epidemiology and public health. *Annals of Epidemiology* 27, 5 (2017), 297–301.

[108] M Sastry and John D Sterman. 1992. Desert island dynamics: an annotated survey of the essential system dynamics literature. In *Proceedings of the 1993 Internatioanl System Dynamics Conference, Cancun, Mexico*. 466–475.

[109] Ellie Senft, Benton Weeks, James Palmer, Benson Neely, Benjamin Turner, and JD Caddell. 2019. A systems dynamics approach to human trafficking in Maharashtra, India. In *2019 IEEE International Systems Conference (SysCon)*. IEEE, 1–7.

[110] Elham Shaabani, Hamidreza Alvari, Paulo Shakarian, and J.E. Kelly Snyder. 2016. MIST: Missing Person Intelligence Synthesis Toolkit. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (Indianapolis, Indiana, USA) *(CIKM '16)*. Association for Computing Machinery, New York, NY, USA, 1843–1867. https://doi.org/10.1145/2983323.2983346

[111] Jia Shao, Bo Du, Chen Wu, and Lefei Zhang. 2019. Tracking objects from satellite videos: A velocity feature based correlation filter. *IEEE Transactions on Geoscience and Remote Sensing* 57, 10 (2019), 7860–7871.

[112] Hamid Reza Sharifzadeh, Hengyi Zhang, and Adams Wai-Kin Kong. 2014. Vein pattern visualization through multiple mapping models and local parameter estimation for forensic investigation. In *2014 22nd International Conference on Pattern Recognition*. IEEE, 160–165.

[113] Jaeho Shin, Christopher Ré, and Michael Cafarella. 2015. Mindtagger: A Demonstration of Data Labeling in Knowledge Base Construction. *Proc. VLDB Endow.* 8, 12 (Aug. 2015), 1920–1923. https://doi.org/10.14778/2824032.2824101

[114] Omer Shwartz, Yael Mathov, Michael Bohadana, Yuval Elovici, and Yossi Oren. 2017. Opening Pandora's box: effective techniques for reverse engineering IoT devices. In *International Conference on Smart Card Research and Advanced Applications*. Springer, 1–21.

[115] 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. In *Proceedings of the 15th Annual International Conference on Digital Government Research* (Aguascalientes, Mexico) *(dg.o '14)*. Association for Computing Machinery, New York, NY, USA, 86–90. https://doi.org/10.1145/2612733.2612746

[116] Nathalie A Smuha. 2020. Beyond a Human Rights-Based Approach to AI Governance: Promise, Pitfalls, Plea. *Philosophy & Technology* (2020), 1–14.

[117] Amazon Staff. 2019. How Amazon Rekognition helps in the fight against some of the worst types of crime. https://www.aboutamazon.com/news/innovation-at-amazon/how-amazon-rekognition-helps-in-the-fight-against-some-of-the-worst-types-of-crime

[118] Jennifer Stoll, W Keith Edwards, and Elizabeth D Mynatt. 2010. Informal interactions in nonprofit networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 533–536.

[119] Jennifer Stoll, W Keith Edwards, and Elizabeth D Mynatt. 2010. Interorganizational coordination and awareness in a nonprofit ecosystem. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. 51–60.

[120] Angelika Strohmayer, Mary Laing, and Rob Comber. 2017. Technologies and social justice outcomes in sex work charities: Fighting stigma, saving lives. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3352–3364.

[121] Abby Stylianou, Abigail Norling-Ruggles, Richard Souvenir, and Robert Pless. 2015. Indexing open imagery to create tools to fight sex trafficking. In *2015 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*. IEEE, 1–6.

[122] A. Stylianou, J. Schreier, R. Souvenir, and R. Pless. 2017. TraffickCam: Crowdsourced and Computer Vision Based Approaches to Fighting Sex Trafficking. In *2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*. 1–8.

[123] Hannah Thinyane and Karthik S. Bhat. 2019. Apprise: Supporting the Critical-Agency of Victims of Human Trafficking in Thailand. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300385

[124] Sabina Tomkins, Golnoosh Farnadi, Brian Amanatullah, Lise Getoor, and Steven Minton. 2018. The impact of environmental stressors on human trafficking. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 507–516.

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

[126] Amna Toor. 2018. Our Identity Is Often What's Triggering Surveillance: How Government Surveillance of# Black-LivesMatter Violates the First Amendments Freedom of Association. *Rutgers Computer & Tech. LJ* 44 (2018), 286.

[127] Edgar Torres, Sergio L Granizo, and Myriam Hernandez-Alvarez. 2019. Gender and Age Classification Based on Human Features to Detect Illicit Activity in Suspicious Sites. In *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, 416–419.

[128] Andrea Tundis, Archit Jain, Gaurav Bhatia, and Max Muhlhauser. 2019. Similarity analysis of criminals on social networks: An example on Twitter. In *2019 28th International Conference on Computer Communication and Networks (ICCCN)*. IEEE, 1–9.

[129] Mary K Twis and Kathleen Preble. 2020. Intersectional standpoint methodology: Toward theory-driven participatory research on human trafficking. *Violence and victims* 35, 3 (2020), 418–439.

[130] UN General Assembly. 2000. Protocol to Prevent, Suppress and Punish Trafficking in Persons, Especially Women and Children, Supplementing the United Nations Convention against Transnational Organized Crime. (2000).

[131] United Nations General Assembly. 2017. Resolution adopted by the General Assembly on 19 December 2016: The right to privacy in the digital age. *United Nations* (2017).

[132] United Nations Office of Drugs and Crime. 2020. *2020 UNODC Global Report on Trafficking in Persons.*

[133] US Department of Justice. 2017. *National Strategy to Combat Human Trafficking.* Technical Report.

[134] Soheil Varastehpour, Hamid Sharifzadeh, Iman Ardekani, and Xavier Francis. 2019. Vein Pattern Visualisation and Feature Extraction using Sparse Auto-Encoder for Forensic Purposes. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 1–8.

[135] Anna Vartapetiance and Lee Gillam. 2014. " Our Little Secret": pinpointing potential predators. *Security Informatics* 3, 1 (2014), 1–19.

[136] Lauren Vollinger. 2021. Concretizing intersectional research methods: Incorporating social justice and action into United States sex trafficking research. *Journal of Human Behavior in the Social Environment* 31, 5 (2021), 599–625.

[137] Emily Wall, Leslie M Blaha, Lyndsey Franklin, and Alex Endert. 2017. Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 104–115.

[138] Emily Wall, John Stasko, and Alex Endert. 2019. Toward a design space for mitigating cognitive bias in vis. In *2019 IEEE Visualization Conference (VIS)*. IEEE, 111–115.

[139] Hao Wang, Congxing Cai, Andrew Philpot, Mark Latonero, Eduard H. Hovy, and Donald Metzler. 2012. Data Integration from Open Internet Sources to Combat Sex Trafficking of Minors. In *Proceedings of the 13th Annual International Conference on Digital Government Research* (College Park, Maryland, USA) *(dg.o '12)*. Association for Computing Machinery, New York, NY, USA, 246–252. https://doi.org/10.1145/2307729.2307769

[140] Bryce Westlake, Martin Bouchard, and Richard Frank. 2012. Comparing methods for detecting child exploitation content online. In *2012 European Intelligence and Security Informatics Conference*. IEEE, 156–163.

[141] Florian Wettstein. 2015. Normativity, ethics, and the UN guiding principles on business and human rights: A critical assessment. *Journal of Human Rights* 14, 2 (2015), 162–182.

[142] Susan Wolfinbarger, Jonathan Drake, and Eric Ashcroft. 2014. Geospatial Technologies and Human Rights Project: Satellite Imagery Assessment of Forced Relocation near Luiswishi Mine. (2014).

[143] Meg Young, Luke Rodriguez, Emily Keller, Feiyang Sun, Boyang Sa, Jan Whittington, and Bill Howe. 2019. Beyond open vs. closed: Balancing individual privacy and public accountability in data sharing. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 191–200.

[144] Aleš Završnik. 2019. Algorithmic justice: Algorithms and big data in criminal justice settings. *European Journal of Criminology* (2019), 1477370819876762.

[145] Ce Zhang, Jaeho Shin, Christopher Ré, Michael Cafarella, and Feng Niu. 2016. Extracting Databases from Dark Data with DeepDive. In *Proceedings of the 2016 International Conference on Management of Data* (San Francisco, California, USA) *(SIGMOD '16)*. Association for Computing Machinery, New York, NY, USA, 847–859. https://doi.org/10.1145/2882903.2904442

[146] Jessica Zhu, Lin Li, and Cara Jones. 2019. Identification and Detection of Human Trafficking Using Language Models. In *2019 European Intelligence and Security Informatics Conference (EISIC)*. IEEE, 24–31.

[147] Michael Zimmer. 2010. "But the data is already public": on the ethics of research in Facebook. *Ethics and information technology* 12, 4 (2010), 313–325.

Received April 2021; revised November 2021; accepted March 2022