The invisible arms race: digital trends in illicit goods trafficking and AI-enabled responses

Preprint · October 2023

DOI: 10.36227/techrxiv.24288703

11 authors, including:

Ioannis Mademlis
Harokopio University of Athens
91 PUBLICATIONS **1,252** CITATIONS
SEE PROFILE

Panagiotis Radoglou Grammatikis
University of Western Macedonia
87 PUBLICATIONS **2,325** CITATIONS
SEE PROFILE

Panagiotis G. Sarigiannidis
University of Western Macedonia
484 PUBLICATIONS **8,462** CITATIONS
SEE PROFILE

Konstantinos Votis
Information Technologies Institute (ITI)
337 PUBLICATIONS **3,833** CITATIONS
SEE PROFILE

The invisible arms race: digital trends in illicit goods trafficking and AI-enabled responses

Ioannis Mademlis, *Senior Member, IEEE*, Marina Mancuso, Caterina Paternoster, Spyridon Evangelatos, Emma Finlay, Joshua Hughes, Panagiotis Radoglou-Grammatikis, *Member, IEEE*, Panagiotis Sarigiannidis, *Member, IEEE*, Georgios Stavropoulos, Konstantinos Votis, and Georgios Th. Papadopoulos, *Member, IEEE*

***Abstract*—Recent trends in the modus operandi of technologically-aware criminal groups engaged in illicit goods trafficking (e.g., firearms, drugs, cultural artifacts, etc.) have given rise to significant security challenges. The use of cryptocurrency-based payments, 3D printing, social media and/or the Dark Web by organized crime leads to transactions beyond the reach of authorities, thus opening up new business opportunities to criminal actors at the expense of the greater societal good and the rule of law. As a result, a lot of scientific effort has been expended on handling these challenges, with Artificial Intelligence (AI) at the forefront of this quest, mostly machine learning and data mining methods that can automate large-scale information analysis. Deep Neural Networks (DNNs) and graph analytics have been employed to automatically monitor and analyze the digital activities of large criminal networks in a data-driven manner. However, such practices unavoidably give rise to ethical and legal issues, which need to be properly considered and addressed. This paper is the first to explore these aspects jointly, without focusing on a particular angle or type of illicit goods trafficking. It emphasizes how advances in AI both allow the authorities to unravel technologically-aware trafficking networks and provide countermeasures against any potential violations of citizens' rights in the name of security.**

***Index Terms*—3D printing, artificial intelligence, AI ethics, cryptocurrency, Dark Web, Deep Neural Networks, graph analytics, security, trafficking, trustworthy AI.**

I. INTRODUCTION

RECENT developments in digital technologies have greatly facilitated trans-border trafficking of illicit goods. During the past decade, it has been made clear that the access to drugs, firearms and other contraband is made significantly easier, at least for people without previous connections to organized crime, by social media, Dark Web marketplaces, cryptocurrency transactions, 3D printers and similarly sophisticated tools [1]. But even organized crime networks and terrorist/extremist groups can benefit from such channels, staying beyond the reach of the law for longer time periods with fewer difficulties.

This situation has increased the stakes for Law Enforcement Agencies (LEAs) and/or public authorities, giving rise to a progressively worsening security environment across the globe. The harmful societal and economic impact of illicit goods trafficking, such as firearms or drugs, is obviously enormous. For example, in 2019 alone, more than 250 thousand people died as a result of firearms worldwide, where nearly 71% of gun deaths were homicides, about 21% were suicides and 8% were unintentional firearms-related accidents [2]. On the other hand, in the European Union alone, over 8,300 deaths involving one or more illicit drugs were reported in 2018 [3], while roughly EUR 3.3 billion are spent on an annual basis for hospital-based drug treatment. More generally, illicit trafficking activities are intricately linked to various forms of violent organized crime and provide funding or manpower for multiple criminal/terrorist groups [4].

Profit- or ideology-driven networks engaging in illicit trafficking are typically interconnected, geographically dispersed, fluid, secretive and flexible in their structure. Thus, the very nature of on-line transactions, coupled with the inherent anonymity offered by cryptocurrency payments and by onion routing-based networking software – like Tor, which is commonly employed for accessing the Dark Web – renders such options attractive to them. As a result, novel criminal modi operandi have emerged during the past decade. In certain cases, traffickers can operate not only through the Dark Web, but also

Date of submission: 15/9/2023. This work received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101073876 (Ceasefire). *Corresponding author: Ioannis Mademlis*.

Ioannis Mademlis is with the Harokopio University of Athens, Department of Informatics and Telematics, Omirou 9, Tavros, Athens Greece, GR17779 (e-mail: yannismademlis@gmail.com).

Marina Mancuso is with the Università Cattolica del Sacro Cuore - Transcrime, Largo Gemelli 1, Milan, Italy, 20123 (e-mail: marina.mancuso@unicatt.it).

Caterina Paternoster is with the Università Cattolica del Sacro Cuore – Transcrime, Largo Gemelli, 1, Milan, Italy, 20123 (e-mail: caterina.paternoster@unicatt.it).

Spyridon Evangelatos is with Netcompany-Intrasoft S.A., 2b, rue Nicolas Bové, Luxembourg, L-1253 (email: spyros.evangelatos@netcompany-intrasoft.com) and with the Hellenic Mediterranean University, 3 Romanou str, Halepa, Chania, Crete, GR73133.

Emma Finlay is with Trilateral Research, Waterford, Ireland, X91W0XW (email: emma.finlay@trilateralresearch.com).

Joshua Hughes is with Trilateral Research, One Knightsbridge Green (5th Floor), London SW1X 7QA (e-mail: joshua.hughes@trilateralresearch.com).

Panagiotis Radoglou-Grammatikis is with the University of Western Macedonia, Active Urban Planning Zone (ZEP), Kozani, Greece, GR50100 (email: pradoglou@uowm.gr).

Panagiotis Sarigiannidis is with the University of Western Macedonia, Active Urban Planning Zone (ZEP), Kozani, Greece, GR50100 (email: psarigiannidis@uowm.gr).

Georgios Stavropoulos is with the Information Technologies Institute/Centre for Research and Technologies Hellas (CERTH/ITI), Thermi, Thessaloniki, Greece, GR57001 (email: stavrop@iti.gr).

Konstantinos Votis is with the Information Technologies Institute/Centre for Research and Technologies Hellas (CERTH/ITI), Thermi, Thessaloniki, Greece, GR57001 (email: kvotis@iti.gr).

Georgios Th. Papadopoulos is with the Harokopio University of Athens, Department of Informatics and Telematics, Omirou 9, Tavros, Athens Greece, GR17779 (e-mail: g.th.papadopoulos@hua.gr).

through regular Surface Web e-commerce sites. For instance, in the case of legally purchased dual-use and (pre-)pre-cursor chemical substances that can be abused by the buyer for producing illegal drugs [5]. A more elaborate practice that is currently becoming more common is to digitally distribute blueprints of desired contraband (e.g., guns), which are then 3D printed using appropriate equipment. In such cases, only specific parts which cannot be reproduced by 3D printing (e.g., metal weapon barrels) need to be physically smuggled through shipping containers, or even via parcels sent through legitimate postal/courier services.

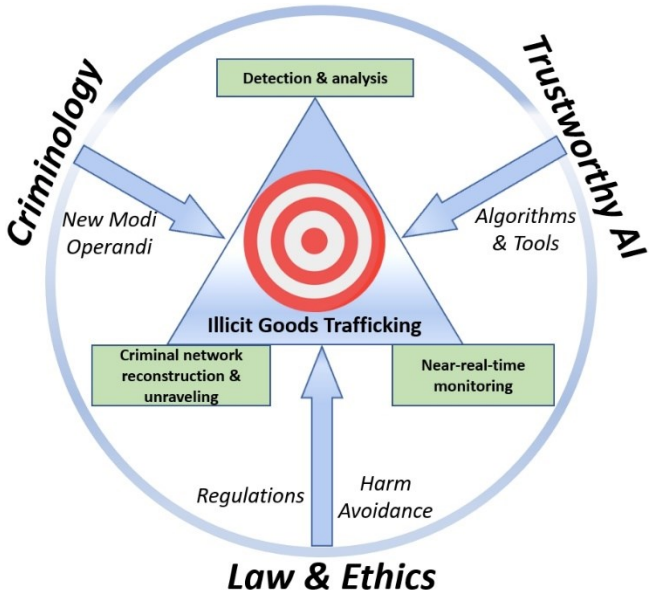

Fig. 1. The conceptual framework of this article entails a synthesis of insights from 3 different disciplines: criminology, AI and legal/ethics studies.

A. Contribution

Advanced Artificial Intelligence (AI) algorithms for large-scale information processing hold the potential to facilitate the detection, analysis, monitoring, reconstruction and unraveling of such technologically-aware trafficking groups, by identifying and correlating their activities and the involved actors. Therefore, AI can serve as a valuable tool for enhancing LEA productivity, response times and ability to counteract the appropriation of high-tech by organized crime. This mostly holds true for modern machine learning and data mining methods, such as Deep Neural Networks (DNNs) and graph analytics, which are able to sift through massive amounts of information in order to automatically detect actionable patterns, insights and trends in minimal time.

However, advanced AI-enabled technology operating on massively collected, large-scale real-world data at the hands of state authorities gives rise to significant ethical and legal issues: the citizens' right to privacy may be placed at risk, inherent bias within machine learning models may reproduce systemic inequalities, while constitutional guarantees may be jeopardized. A potential technical solution to these concerns arises from the fledgling scientific subfield of Trustworthy AI [6], [7], which promises to mitigate similar dangers by imparting AI algorithms with inherent robustness against them.

Overall, this paper explores the above issues synergistically, emphasizing how advances in AI *both* allow the authorities to unravel technologically-aware trafficking networks *and* provide countermeasures against any potential violations of citizen rights in the name of security. To the best of the authors' knowledge, no previous recent article either investigates this domain from such a combined perspective, or treats all illicit goods trafficking activities at once. Existing papers tend to focus only on a specific type of illicit trafficking and its particularities [8], [9], do not always engage with modern digital trends employed by traffickers and deal with either exclusively technical [10], [11], exclusively criminological [12] or exclusively ethical/legal issues [13], [14]. The conceptual framework of this article is graphically summarized in Fig. 1.

II. METHODOLOGY

The scope of this article covers rapidly evolving themes at the intersection of multiple disciplines. Therefore, a scoping review methodology was employed, which aims to map the available literature, identify key concepts and trends, as well as clarify gaps in a selective manner. This approach allows exploration and synthesis of diverse research findings across multiple fields. The search strategy involved querying multiple academic databases, including Google Scholar, IEEE Xplore, and Springer, using targeted relevant keywords. The "Related Work" sections of each paper were also utilized, to identify additional relevant studies. To ensure contemporary relevance, the search primarily focused on publications from the last three years, for very active sub-fields, or the last seven years, for less active subfields. Seminal works were also included, when deemed significant. Initial screening was based on titles and abstracts, followed by full-text review of selected papers. To maintain focus, avoid redundancy, and respect article length limitations, only 1-2 representative papers were typically included from each set of similar studies, based on relevance and impact.

The collected articles on state-of-the-art AI methods for large-scale information processing were subsequently thematically clustered, according to how they are applied to the examined application domain. In general, such methods can be leveraged for detecting, analyzing, monitoring, reconstructing and unraveling technologically-aware trafficking networks, which follow the trends to be presented in Section III. The methods were grouped into two primary categories: a) Deep Neural Networks (DNNs), and b) Graph analytics. Graph Neural Networks (GNNs) essentially form a hybrid approach, being neural networks that allow knowledge mining on graphs. After clustering, the collected articles were further assessed for their ability to process large-scale data and their relevance to law enforcement applications. Additionally, real-world use of such methods gives rise to ethical and legal concerns. These were first identified through a detailed review of legal and ethical frameworks, then subsequently clustered into three key areas: human dignity and autonomy, bias and discrimination, and privacy and data protection. This

clustering informed the discussion of Trustworthy AI solutions, which align technical innovations with legal and ethical compliance.

The above methodological approach was designed based on the fact that this study was guided by the following two research questions:

R1: How can advanced AI methods be applied to detect, analyze, monitor, reconstruct, and unravel technologically-aware trafficking networks? This question seeks to explore the practical applications of AI technologies, focusing on their ability to process large-scale information in domains such as image analysis, social network mapping, cryptocurrency transaction tracing, and on-line communication monitoring.

R2: What are the key ethical and legal challenges associated with deploying these AI methods in law enforcement, and how can they be addressed through Trustworthy AI principles? This question examines the societal implications of using AI for combating illicit goods trafficking, focusing on issues like bias, privacy, human rights, and compliance with emerging regulations.

Jointly, these questions provided a focused framework for the article's discussion and structure, ensuring that the exploration of AI methodologies and their implications is both targeted and impactful. These questions are explored through a thematic review of AI methodologies in Sections IV and V and a detailed examination of ethical and legal issues in Section VI. The above-described methodological approach not only maps current capabilities and gaps, but also identifies pathways for aligning AI applications with law enforcement needs and societal expectations.

A. Outline

The remainder of this article is structured in the following manner. Section III presents and discusses recent criminal modi operandi that have emerged in illicit goods trafficking, thanks to modern digital technologies. Sections IV and V briefly overview state-of-the-art AI methods that can facilitate the detection, analysis, monitoring, reconstruction and unraveling of technologically-aware trafficking networks. Section IV focuses on DNNs, while Section V on graph analytics. This discussion is *not* deeply technical and emphasizes how AI algorithms for large-scale information processing can be employed to assist LEAs. Section VI reviews legal and ethical concerns that arise from the extensive employment of AI by LEAs, along with potential mitigation countermeasures offered by Trustworthy AI methods. Finally, Section VII draws conclusions from the preceding discussion.

III. EMERGING CRIMINAL MODI OPERANDI EXPLOITING DIGITAL TECHNOLOGIES

The advent of modern digital technologies has revolutionized the way in which certain traditional crimes are perpetrated. For example, the World Wide Web, with its three main layers (Surface Web, Deep Web and Dark Web) has given rise to entirely new forms of criminal activity (e.g., phishing, the use of malware or ransomware, etc.), but it has also altered the *modi operandi* used to commit crimes. It has multiplied the avenues for financial fraud, the access to vulnerable children by child-sex offenders, and the possibilities for illicitly trafficking goods such as drugs and firearms [15].

A. The Dark Web

The Dark Web is a part of the Deep Web accessible only through specific browsers, usually Tor, which ensures anonymity to the users by essentially hiding their IP address from surveillance and traffic analysis while navigating. The anonymity provided by the Dark Web makes it attractive to individuals interested in carrying out criminal activities with a low risk of being tracked and identified, including the sale and purchase of illegal goods [16], [17], [18].

Within the Dark Web, two types of marketplaces are employed to traffic illicit goods: vendor shops and cryptomarkets. Also known as "single-vendor markets", vendor shops are on-line stores administered by a single seller, who directly manages the transactions with customers, without third-party services. The direct connection between seller and buyer decreases commissions' costs and risks associated with third-party services [19], [20]. According to Europol [21], in recent years there has been a rise in single-vendor shops in the Dark Web, which frequently rely also on encrypted communication platforms, such as Wickr and Telegram, to manage sales. These findings highlight a tendency in criminal activities towards more decentralized and secure channels that establish higher trust among involved parties.

B. The Surface Web

On the other hand, similarly to legal e-commerce platforms on the Surface Web, cryptomarkets offer the following possibilities: a) for multiple vendors to display their goods available for sale by posting images and descriptions of the goods, and b) for buyers to search for a specific good. Cryptomarkets provide more visibility in comparison to single-vendor shops, offering access to a broader client base for sellers and third-party services to make transactions safer for buyers. An example is the escrow, which ensures that the cryptocurrency payment is held by the market until the buyer confirms reception of the goods, or a designated waiting period has elapsed, at which point the transaction is made accessible to the vendor. Additionally, this way the seller preserves his/her anonymity and minimizes the chance of the illicit transaction being traced back to him/her [19], [22], [16], [23], [24], [25], [18]. The quality and the assurance of delivery are then assessed by buyers, who are thus able to impact vendors' reputation [26].

Dark Web markets, and more generally the Internet, have allowed for a relevant innovation in illicit goods trafficking by decreasing the effort required to conduct it [27], [28]. Illicit goods, such as drugs and firearms, can be ordered and paid

just by clicking a button, while being received in a few days across the world. Buyers and sellers do not need to meet at a physical location; they can conclude more illicit transactions via the Internet, in a quicker and safer manner and without sharing their identities or contact information [25], [18]. The on-line availability of illicit goods allows an easier matching of demand and supply in the case of inexperienced individuals with no direct or indirect criminal connections, which is instead a fundamental feature of off-line sales [29], [30].

Goods sold on-line are delivered mainly through postal and courier services, which allow vendors to easily and quickly send goods anywhere in the world. Shipments in this case are quite small (multiple small shipments can be sent at a time) and the addresses of the recipients do not correspond to personal addresses of the buyers [31]. Other methods used include concealing goods inside legitimate items and making the parcel appear professional [31], [32]. Indeed, on-line marketplace forums are used by vendors to discuss shipping options for avoiding law enforcement detection, based on criteria for profiling suspect packages that are regularly published by governments [19]. Nonetheless, in recent years, vendors and buyers have started avoiding traditional courier shipments when possible, relying on the so-called "dead drop" method. It involves paid intermediaries who first conceal the prepackaged illicit shipment in discreet locations, and then share with the vendor the coordinates through a short video for each "dropped" deal. The vendor then sends the geo-coordinates to the client, who can eventually pick up the goods [23], [19].

It is estimated that approximately 62% of the illicit goods sold in Dark Web markets are drugs and drug-related chemicals [33]. The on-line trade of other contraband, such as firearms, remains limited in volume and in value [34]. In fact, firearms and explosives account for only about 1% of illicit goods smuggled on the Dark Web [33]. Over the years, the sale of these goods has shifted from Dark Web marketplaces to Surface Web forums and secure platforms, such as Telegram and Wickr, following prohibitions on firearms sales in several marketplaces [34], [21]. Within the European Union, relevant discussions on on-line forums and social media, such as Twitter, YouTube, Reddit, and Discord, mainly concern guidelines and instructions for manufacturing firearms at home, especially with the use of 3D printers by "enthusiasts"; this is an emerging modus operandi [35], [19]. Such digital products, also available in Dark Web markets as blueprint files for sale [21], are associated with even lower risks due to the lack of any physical exchange [19].

C. Gaining Access to Illicit Goods and Services

In general, on-line avenues tend to provide access to illicit items to prospective buyers that do not otherwise have connections to traffickers, while the organized criminal networks still strongly prefer off-line channels when acting as purchasers. However, despite their currently smaller size compared to the traditional off-line market, Surface and Dark Web may become a more important marketplace in the near future for the trafficking of illicit goods other than drugs. The use of on-line channels increased a lot, for example, due to the restrictions caused by the Covid-19 pandemic, which led to a re-orientation of logistical chains. However, scarce information is available on the long-term impact of the pandemic on trafficking patterns [36].

Illicit drugs trafficking is clearly the domain most affected by digital technologies, with the variety of on-line exchange avenues being greater than in the case of other types of contraband. On the Dark Web, drug sales can occur within a marketplace, within a decentralized network or between individuals [37]. The most traditional recreational drugs such as cannabis, MDMA and LSD, as well as certain prescribed medicines, are the most popular in marketplaces [38]. Sales of new psychoactive substances and legal pharmaceutical products are traded mainly on the Surface Web [37]. Overall, four primary avenues have been identified: a) e-shops selling new psychoactive substances as research chemicals, mostly under their chemical names; b) e-shops selling products under brand names; c) classified ads, often located within public Web sites; and d) a Deep Web avenue [39]. These routes suggest a growing hybridization between the commercial and research chemical areas and the presence of a "grey market", which includes Web sites having both a Surface Web presence and a hidden Deep Web element [37].

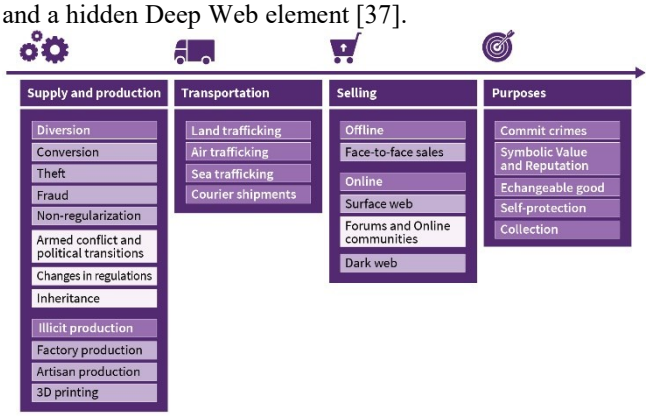

Fig. 2. A visual summary of the current criminal modi operandi in illicit goods trafficking.

D. The Use of Cryptomarkets

The use of cryptomarkets also increased during the Covid-19 pandemic [40]; the physical lockdown measures pushed many people to order drugs on-line. However, the restrictions imposed by many governments may have negatively impacted the function of these markets as well, especially concerning international sales, since drug dealers mainly hide illicit drug loads within legitimate international shipments [22]. In fact, the number of unsuccessful transactions on cryptomarkets increased during the pandemic due to delivery failures, related to the international/inter-continental nature of the transactions and the severity of the crisis in the vendor's country [41]. Despite this, the opportunities offered by the Internet continue to foster the proliferation of the on-line drug markets. The closure of major Dark Web-hosted platforms seems to have a minimal long-term impact, since vendors can easily migrate to other platforms [42].

Fig. 2 summarizes the current state of the criminal modi

This article has been accepted for publication in IEEE Transactions on Technology and Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TTS.2024.3514683

operandi in illicit goods trafficking, emphasizing contemporary digital trends.

IV. FIGHTING ILLICIT GOODS TRAFFICKING USING DEEP NEURAL NETWORKS

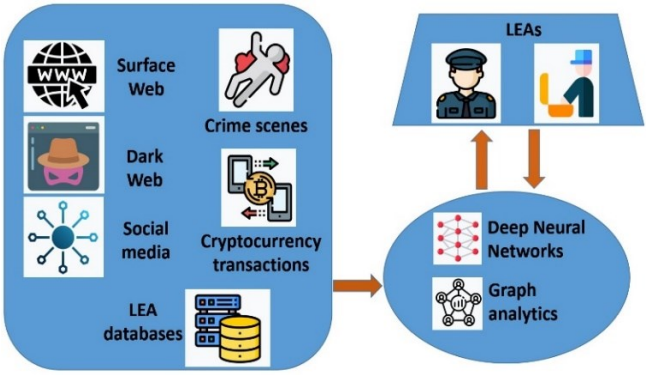

Fig. 3. High-level functional diagram of AI deployment for fighting illicit goods trafficking.

Both DNNs and graph analytics can be employed for fighting modern illicit goods trafficking. Fig. 3 depicts an example functional diagram, visualizing the relation between these methods, LEA users and the required data sources. Below, this section details the use of DNNs in this context, while Section V focuses on the use of graph analytics.

DNNs are complex machine learning algorithms that mimic the human brain at a very abstract level and in a very simplified form. Thus, a DNN is composed of many simple, interacting and interlinked computational units called *neurons* that are arranged in consecutive *layers*. It is typically trained to learn a desired mapping from input data samples (e.g., images, text, videos, etc.) to respective outputs (e.g., classification predictions). DNNs are defined by a high number of parameters (e.g., in the range of millions or billions) that are being learned automatically by a training algorithm. Training is usually performed on a large, annotated dataset of known input data samples (most typically in the range of thousands or tens of thousands, but even larger datasets exist), for which the desired corresponding outputs have been prespecified by human annotators. The process results in specific parameter values, which jointly define a particular trained DNN model.

These parameters get frozen when training finishes and, from then on, the trained DNN model is ready to analyze novel/unknown test input data at the so-called *test stage*, to make useful predictions about them. Essentially, this is the actual operational deployment phase for the DNN. In case the test data are, at a certain point, noticed to be significantly different from the training dataset (e.g., due to real-world variations accumulating over time), then the model has to be temporarily withdrawn and retrained with a new annotated dataset that better reflects the current situation.

The majority of downstream AI tasks are either classification tasks or regression tasks, with certain ones (e.g., object detection on images) combining aspects of both. In classification, the goal is to predict for each input a specific class label, among a set of K predefined classes (e.g., "firearms", "drugs" or "non-suspicious"). Instead, in regression the predicted output is one or more real numbers. Hybrid tasks are characterized by domain-specific outputs; for instance, in object detection on images the output is a set of image locations where objects-of-interest are depicted, along with the corresponding class labels.

A. Advantages of DNNs over Traditional ML Methods

DNNs have a definite edge over traditional machine learning methods, thanks to their ability to automatically learn optimal, internal numerical representations of the input data, which eliminates the need for manual feature extraction using human-made and domain-specific algorithms. Different neural architectures have been proposed over the years, typically varying with regard to how their layers operate. Thus, Convolutional Neural Networks (CNNs) are mostly suitable for image analysis, since they receive their input in the form of tensors instead of vectors. Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks are mostly employed for analyzing time series and other types of sequences, thanks to their limited internal memory. Currently, Transformer architectures relying on the so-called *self-attention* mechanism empower the state-of-the-art algorithms in most AI tasks. Their strength lies in their ability to efficiently capture long-range dependencies and contextual information within sequential data of any nature (e.g., textual documents, videos, etc.), or even images.

B. Diverse Data Sources for DNNs

DNNs have proven extremely effective in perception and language-analysis tasks, particularly if there is a high volume of data to be analyzed. Naturally, their usual role with regard to fighting illicit goods trafficking falls under this area. The majority of relevant work exploits DNNs for object detection in image/video data, object recognition in images and for Natural Language Processing (NLP). The relevant data are typically massive in volume and gathered from the following sources:

- **Surface Web**. Web sites/fora, e-commerce sites. Typically, the goal of analyzing data from the Surface Web is to detect/monitor legitimate e-shops selling products that may facilitate illicit goods trafficking (e.g., licit drug precursors), identify discussions among users related to trafficking networks, or even discover blueprints for 3D printing of illicit goods (e.g., guns).
- **Dark Web**. Dark Web marketplaces. Typically, the goal of analyzing data from the Dark Web (e.g., images, text, etc.) is to directly identify illicit marketplaces where trafficking networks may thrive. Of course, given that the Dark Web is by definition concealed, it is not always easy to access relevant data and collect them for AI-enabled analysis, but dedicated Dark Web crawlers facilitate the process [43]. In a complementary fashion, Darknet network traffic data can

also be collected and analyzed, even allowing real-time monitoring.

- **Social media**. Social media platforms, messaging apps. These are typically analyzed in order to identify discussions about and/or covert marketing/advertising of illicit goods, while also flagging users with an active interest in them. Text and images are the usual data modalities extracted from such sources, with facial recognition of suspects using photographs posted in social media (e.g., through the proprietary *Clearview* software) becoming more common during the past decade [44].
- **Crime scenes or entry points**. These are typically images/videos, along with relevant metadata, that are physically captured at crime scenes (e.g., at drug or gun shipment seizures) or on potential entry points for traffickers (e.g., customs offices, ship ports, airports, etc.). In the case of entry points, relevant data may have been captured en masse for prevention purposes (e.g., CCTV monitoring, X-ray scanning of all mailed parcels, millimeter-wave imaging of passengers, etc.). In rarer cases, data modalities may extend beyond the "visual + metadata" domain; for instance, chemical signatures derived from drug screening kits, results from vapor-based inspection of maritime containers, toxicology reports related to drug overdoses, etc.
- **LEA databases**. These are typically structured or unstructured data maintained by LEAs, both historical and current. They may contain almost any type of modality that can be analyzed by DNNs (text, audio, visual, etc.), potentially along with relevant metadata. The original data sources may also vary greatly, ranging from targeted wiretaps to massively crawled social media posts.

In the remainder of this section, an overview of the most important aspects of the relevant state-of-the-art is presented, organized per data modality.

C. Visual Data Analysis

Regarding the fight against illicit goods trafficking, the typical visual data of interest are RGB images, RGB videos, and X-ray scans or millimeter-wave images. DNNs are naturally at the forefront of relevant automated analysis. Regular RGB image analysis has been extended to the case of illicit goods trafficking, with the vast majority of literature focusing on object recognition or object detection [45]. Since there is a considerable domain and task overlap with public safety applications (e.g., real-time monitoring of public spaces to prevent terrorist attacks, surveillance of general criminal activities, etc.), the volume of related research is large. However, the dominant trend is to simply adapt popular DNN solutions for generic image analysis, using image datasets depicting illicit goods. The application-specific challenges are mostly three: a) scarcity of very large, annotated relevant datasets, b) small on-image size of many objects-of-interest, and c) large variation in viewpoints, background clutter and levels of visual occlusion between different images depicting the same type of illicit good [46]. Additionally, unlike public safety monitoring, there is rarely an actual need for real-time processing. Most recent work of this kind employs CNNs for whole-image classification [47], [48] or for object detection purposes [49], [50], [51]. Well-known relevant neural architectures are typically utilized and properly adapted, such as VGG-16, YOLOv3-5, SSD, Faster R-CNN, with the choice mainly dictated by the desired inference speed-accuracy trade-off and the particularities of the specific dataset employed. In general, less complex architectures are faster and less prone to overfitting when trained with comparatively small datasets, but more complex DNNs may achieve greater test-stage accuracy when trained on sufficiently large datasets.

Very few recent approaches deviate from the paradigm described above; this also holds true for RGB video analysis (mainly CCTV footage), where video frames are simply extracted and analyzed as separate images [52]. At most, each video frame may first be preprocessed to highlight any human bodies visible in the image, using a separate, pretrained DNN for person detection [53] or for human body pose estimation [54], [55]. A notable example that deviates from the norm is the method in [56], which exploits an ensemble of multiple, simple CNNs, instead of one monolithic DNN, for weapon detection in images. Each model detects a specific firearm component/part (e.g., barrel) and their outputs are aggregated to obtain a single final prediction.

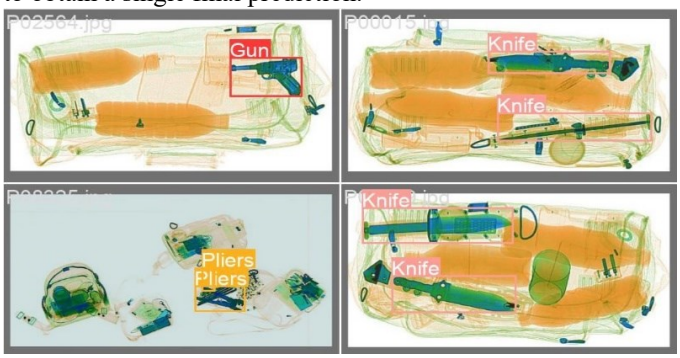

Fig. 4. An example of illicit item detection on an X-ray scan of a passenger luggage, using DNNs [45].

DNNs also dominate modern scan image analysis for the X-ray and millimeter-wave modalities. The typical goal is to automatically detect illicit goods, such as drugs or weapons, in passengers, luggage or mailed parcels; an example is shown Fig. 4. The dominant trends are similar to those of the RGB image analysis (e.g., [57], [58]), but obviously different training datasets are utilized. However, special neural modules are commonly employed as part of the overall DNN [59], [60], [61], [62], so that accuracy is improved in the face of typical issues such as occlusions, cluttered background or class imbalance. Finally, a few more idiosyncratic attempts have been made towards exploiting visual analysis DNNs to more explicitly identify on-line criminal actors. For instance, the method in [63] exploits CNNs for illicit vendor re-identification/fingerprinting in the Dark Web, through recognizing the style of their posted photographs.

D. Language/Text Analysis

NLP algorithms are mainly utilized to identify discussions, user/customer reviews or advertisements concerning illicit goods, or facilitators for their production (e.g., drug precursors, 3D printing blueprints of firearms). The data sources can be any type of on-line content, text from LEA databases or transcriptions of audio stored in LEA databases (e.g., from wiretaps). The vast majority of relevant literature concerns classification of short texts as being about specific types of illegal activities (e.g., [64] for detecting drug dealers).

DNNs dominate the state-of-the-art, as in the majority of NLP tasks. The current trend is to use Transformer architectures that have been pretrained as language models on huge text corpora without human annotations (e.g., BERT, or the GPT family of models), for extracting semantically meaningful word/sentence/document representations in the form of dense numerical vectors. These representations of the text in question are either fed to an appended neural classification head, or are being analyzed by an independent neural network which has been separately trained for classification (e.g., an LSTM, or even a simple MLP). Representative examples include [65], for detecting human trafficking through on-line customer reviews, and [66], which extends this concept for analyzing hashtags in social media posts, in order to identify on-line drug dealers. An alternative approach in [67] exploits multimodal fusion, by proposing a DNN which combines and analyzes text encodings from BERT and image representations from a pretrained CNN, so as to recognize drug dealers operating in social media.

Besides text classification, Named Entity Recognition (NER) is also very prominent in the fight against illicit goods trafficking. It is a separate NLP task that can be either employed as a preprocessing step, or used on its own. It concerns the automatic identification of novel words as referring to named entities (persons, locations, product brands, etc.). The relevant state-of-the-art relies on neural networks [68], but auxiliary exploitation of external knowledge bases (e.g., Wikipedia) via entity linking is also common [69]. The NER outputs can subsequently be employed for identifying criminal events, modeling criminal groups and discovering the structure of trafficking networks.

A widespread issue in NLP technologies is the scarcity of large datasets for training appropriate DNNs in most languages other than English. Currently, the common solution employed when this problem arises is to utilize multilingual automated translation technologies – also powered by DNNs – as a preprocessing step, in order to convert all inputs to English before analyzing them. Alternatively, cross-lingual NLP can handle multiple languages and transfer knowledge from high-resource languages (like English) to low-resource ones [70].

E. Network Traffic Analysis

A dedicated body of research concerns Darknet network traffic analysis. This typically means recognizing Tor network packet exchanges and/or classifying them with regard to the application/service type. This is a rather difficult endeavor, due to the heavy encryption employed by software such as Tor. The state-of-the-art approaches employ DNNs to achieve this, by analyzing both communication patterns/metadata and the encrypted payload. Examples include [71] and [72], where custom combinations of CNNs and LSTMs are proposed, or [73], where a sophisticated CNN architecture is utilized. In this vein of work, the neural (self-)attention mechanism has been indicated as critical for achieving good accuracy. Thus, in [74], a Transformer architecture has been successfully utilized for detecting Tor traffic, after training on an artificially augmented dataset (using synthetic data). Alternatively, in [75], a Transformer is combined with an LSTM before the classification stage, so that both global (session-level) and temporal (individual packet-level) features are captured. Such methods can be combined with more direct visual processing [76] or NLP [77], so that specifically illegal activities can be identified. Recent relevant approaches may utilize graph structures (see Section V) that capture client-server timestamped interaction patterns within the Darknet, in order to analyze them via GNNs and/or attention mechanisms. This results in improved classification accuracy, by transforming the traffic classification problem into a graph classification one [78]. A systematic review of the field can be found at [79].

V. FIGHTING ILLICIT GOODS TRAFFICKING USING GRAPH ANALYTICS

A graph is a mathematical structure composed of nodes (also called vertices), which typically represent entities, and edges – potentially weighted and/or directed – that connect the nodes according to a certain pattern. Graph analytics/mining algorithms form the heavy AI machinery when it comes to discovering and identifying trafficking networks that follow digital trends. The relevant graphs mostly fall into three categories:

- **Social network graphs**: These graphs represent relationships between individuals, groups or organizations (nodes) in a social context. They focus on capturing social interactions, friendships, collaborations, hierarchies and other social ties (edges). They can be constructed based on data from on-line social media platforms, known or suspected real-world social relationships between persons, or both.
- **Cryptocurrency transactional graphs**: These are graphical representations of the transactions that occur within a cryptocurrency network, such as Bitcoin, Ethereum, Monero, etc. These graphs capture the flow of funds (edges) between different cryptocurrency addresses/wallets (nodes) and provide insights into the transactional behavior within the network. They can be constructed based on the publicly available transaction data stored in the cryptocurrency's blockchain, i.e., a decentralized, encrypted and transparent data structure that records all transactions within the cryptocurrency

network.

- **Knowledge graphs**: These graphs represent structured factual knowledge about entities and their relationships in a specific domain. In general, they can be considered to arise from a hybridization of graph theory with symbolic AI. Nodes express entities (e.g., people, places, events or concepts) and edges logical connections between them. Entities and connections are typically instantiations of a predefined domain-specific ontology, which permits automatic formal reasoning to be applied on the graph once it has been defined. Knowledge graphs can be populated based on inputs from various and diverse sources, such as the outcomes of the DNNs presented in Section IV, outcomes of social network graph or cryptocurrency transactional graph analysis, other automated processing and/or manually entered information (e.g., by LEA officers).

Several different types of graph analytics/mining algorithms can be applied to these three graph types. The main relevant tasks are briefly reviewed below, organized per graph type.

A. Social Network Graphs

All forms of organizations, including criminal ones, are based on human interactions and relations. For this reason, they can be conceptualized as networks, which consist of a set of actors and the relations between them [80], [81]. In the last decades, the network concept and the techniques associated with this approach have been employed in in various way for the study of criminal groups, the trafficking they are involved in, and the possible strategies for their disruption [82].

The most common task related to combatting illicit goods trafficking is "node classification". For instance, the method in [83] trains a GNN to classify social media users as drug dealers or not, based on their on-line relationships and the content they post on-line. Similar approaches could be applied to real-world social networks, as well. These can be modelled using gathered intelligence and crime script analysis methods, which aid in identifying criminal modi operandi, recognizing specialized skills and mapping organizational structures [84]. More often, though, real-world criminal social networks are reconstructed semi-automatically by LEAs through phone communication records. Thus, the method in [85] applies GNNs to such a network.

An equally important task is "community detection", i.e., identifying groups or communities of closely connected nodes within a criminal social network. Uncovering clusters of individuals or entities with strong interactions is a cornerstone of unraveling criminal networks. Classical graph partitioning algorithms based on minimum-cut can be employed: such methods operate by identifying how a graph can be partitioned into disjoint subsets of nodes, in a way that minimizes a chosen metric. However, modularity maximization alternatives, such as the Louvain method [86], are among the most popular choices (e.g., in [85]). Graph-theoretic centrality measures, which quantify the importance of nodes within a graph based on their connectivity, may not only identify nodes that play significant roles in the network, but can also be employed for community detection; this is how the popular Girvan–Newman algorithm operates [87]. Thus, the method in [88] exploits centrality measures within a custom graph clustering algorithm, for identifying communities of criminal actors in social media that engage in human, firearms and drugs trafficking. However, node "influence estimation" is an important task in itself. For instance, centrality measures have been employed for recognizing influential suspicious domains in the Darknet, using graphs where the nodes are Tor hidden services [89].

The task of "link prediction" is particularly significant in criminal social network analysis, since a portion of critical graph connectivity information is typically missing/unknown; even intentionally hidden by traffickers to obstruct LEA investigations. Link prediction algorithms identify missing or future edges and, therefore, can be used to propose potential relationships between suspects or criminal actors, which have not been explicitly observed. For instance, the method in [90] proposes a simple link prediction algorithm, as a preprocessing step for improving community detection. Similarly, the survey in [91] concludes that the most robust link predictor for organized criminal networks is the Katz index, which assigns a score to potential links based on the number of paths that connect them and the lengths of those paths. However, methods based on DNNs tend to become the norm in recent years, such as GNNs [92] or Deep Reinforcement Learning [93].

B. Cryptocurrency Transactional Graphs

The financial transactions in cybercriminal contexts have grown in the last years. The use of cryptocurrencies is often associated to the buying of illicit goods online, even though the cryptocurrency market is now partially regulated [94]. Most of the existing studies lacked detailed accounts of how payments or transactions were conducted or were based on anecdotal cases, limiting the generalizability of their results. Additionally, offenders varied in their use of virtual currencies. Some relied on a single wallet for all their illicit transactions, while others used different wallets for each transaction to maximize their security [94].

Thanks to the transparency offered by blockchain technology, considerable work has already been done on tracing cryptocurrency transactions and analyzing their patterns. However, despite recent improvements in tracing algorithms, they still generate a vast amount of data, with only a few relevant or interesting data points for investigators [94], [95]. Blockchain allows the construction of an evolving graph, where user addresses form the nodes and fund flows are the edges. These graphs exhibit small-world properties [96], meaning that most nodes are not directly connected but can be reached from any other node through just a few steps, thanks to a few highly connected vertices. Analyzing them is extremely important for combatting digitally-aware illicit trafficking groups that operate via the Dark Web. Most tasks that are described in Section V.A in relation to social networks

are also valid in the case of cryptocurrency transactional graphs as well, with several variations of relevant well-known algorithms having been developed for this scenario. An example would be to use GNNs for classifying nodes, as involved in illicit transactions or not [97], [98]. Optionally, simple, handcrafted rule systems can then be developed, by exploiting the domain knowledge of LEA experts. These rules sets will provide desired predictions based on the outputs of the machine learning modules. Such a pipeline is depicted in Fig. 5.

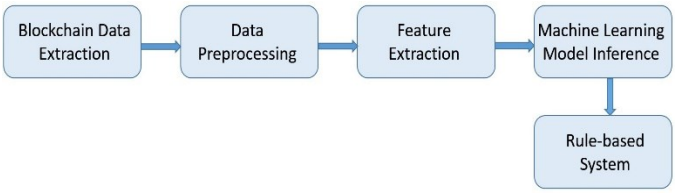

Fig. 5. An example pipeline for automated analysis of cryptocurrency transactional graphs.

However, more specialized methods also exist for tracing cryptocurrency payment flows, i.e., the main interest of LEAs. In criminal investigations, these can be employed to circumvent the anonymity provisions offered by Bitcoin and many later cryptocurrencies. Two such tasks are the most important ones: "address clustering", which refers to identifying a cluster of different addresses/wallets that can be linked with a single human actor, and "attribution tagging", i.e., discovering context information and assigning it to an address or transaction. Tags can flag addresses associated with illegal activities, known criminals, or suspicious behavior. Address clustering and attribution tagging operate synergistically. For example, a single tag attributed to a wallet address may lead to its real-world identification (e.g., IP address) and, consequently, effectively de-anonymize an entire cluster of wallets.

Various heuristics have been proposed for address clustering over the years, with the most prominent one being the Multiple-Input Heuristic (MIH): if two addresses serve as inputs in one transaction, and one of these addresses along with an independent third one serves as inputs in a separate transaction, then most likely the three addresses are controlled by a single individual, who possesses the private keys for all of them [99]. However, besides such heuristic rules, typical node clustering algorithms relying on machine learning have also been applied for address clustering [100], while similar approaches have been combined with anomaly detection to automatically flag potentially illicit transactions [101].

Such methods have been successfully employed to trace suspicious monetary flows in cryptocurrency funds. For instance, a criminal network of human traffickers has been identified by correlating specific Bitcoin transactions with sex ads [102]. As a result, novel "mixing services", which obfuscate fund flows by mixing together irrelevant transactions (e.g., the CoinJoin mechanism), and alternative cryptocurrencies with enhanced privacy support have surfaced in recent years, becoming popular with criminals [103]. State-of-the-art research has been attempting to combat this phenomenon by supplementing graph mining with network analysis methods [104], [105] or with additional features to enhance clustering in the presence of CoinJoin transactions [106].

C. Knowledge Graphs

A knowledge graph is constructed in two steps: first a domain-specific ontology is defined and, subsequently, the graph is populated based on it, using existing structured knowledge bases (e.g., LEA databases) and the outcomes of entity/relationship extraction from unstructured data. In the case of fighting illicit goods trafficking networks, the ontology typically has to be defined in collaboration with criminologists and LEAs. The methods presented in Sections IV, V.A and V.B may serve as automatic mechanisms for extracting entities and relationships from massive volumes of data.

The main benefit of a structured knowledge graph is that it facilitates automatic query answering and formal reasoning, which permit automatic information retrieval and inference of new knowledge. The disparate sources which were previously exploited to populate the graph are implicitly fused in the process, which allows also the automatic detection of conflicts between different sources [107]. Moreover, different knowledge graphs constructed from heterogeneous sources can be merged via algorithms for entity and ontology alignment [107]. Overall, these functionalities can be used for a variety of purposes: automatically linking digital traces (e.g., on-line activities or communication patterns) to physical identities or potential suspects, providing LEA investigators with actionable intelligence (e.g., structure of the trafficking network, key actors, modus operandi, or potential vulnerabilities), predictive analytics (e.g., anticipation of criminal behavior, identification of potential threats), criminal profiling and risk assessment of suspects, decision support in investigations, as well as visualization of trafficking network structure (clusters/communities, central actors, hierarchies).

This ambitious vision has already been partially materialized in relevant research, but typically exploiting only a limited number of data sources and modalities. Thus, the method in [108] exploits a custom ontology and a knowledge graph formalism for facilitating criminal investigations and automating the generation of digital evidence admissible in court, while taking care of the chain of custody. However, it focuses exclusively on social media users and exploits only content automatically crawled from such platforms. A partially similar system is presented in [109], where the knowledge graph is populated by automatically crawling and analyzing on-line newspapers, instead of social media content. In a parallel direction, the system in [110] populates the knowledge graph only via semi-automatic Surface Web crawling, while allowing the non-technical user to define/personalize the domain (e.g., illicit firearms trafficking), the ontology, the crawling keywords and the semantic queries. On the other hand, the method in [111]

implements a different trade-off: it exploits more varying data sources, such as official crime reports, socioeconomic information and geographic information, which are manually entered into an RDF knowledge graph, instead of automatically crawled, in order to facilitate the discovery of criminal patterns (exclusively for Mexico City). Other approaches still try to exploit even more heterogeneous data sources, but limit their scope to specific criminal cases [112].

Among the most complete related research prototypes are the ones presented in [113] and [114]. The first one populates the knowledge graph via automatic analysis of the Surface Web, the Dark Web and cryptocurrency transactions, while it has been shown to be able to identify connections between illegal trafficking of different types of goods. The second one is even more heterogeneous: it automatically collects information from the Surface Web, the Dark Web, social networks, financial data, road traffic data, geospatial data, etc.

All the AI methods for fighting illicit good trafficking presented in Sections IV and V could provide LEAs with useful tools to prevent and fight again this crime. These methods and tools are developed by researchers, who should gather from end-users specific requirements and needs according to their operational experience. The developed tools should reflect these needs and be developed having in mind their final usage. Ad hoc training sessions and testing phases are fundamental in order to ensure end-users will be concretely able to adopt these tools in their day-by-day activities and take advantage from them.

VI. DISCUSSION: ETHICAL AND LEGAL ISSUES

AI-powered law enforcement relying on the methods presented in Sections IV and V has the potential to create a wide variety of tangible and intangible (unintended) harms. In this section, we focus our analysis on individuals who could be affected by use of AI, but it is important to note that other entities could be affected. For example, bias in AI systems might not just affect the individual, but also the reliability of court decisions [115] or the treatment of affected societal groups [116]. Since these systems will process the personal data of members of the public who are innocent until proven guilty, it is important that potential harms and benefits are understood. It is therefore crucial that ethical considerations, legal compliance, and societal acceptability are considered together during the design process of similar AI systems, along with harm-avoidance and benefit-seizing measures. This section first reviews the relevant ethical and legal issues, before proposing the use of "Trustworthy AI" to overcome a number of them through technical means. Where it is necessary to ground concerns to specific regulatory frameworks, EU legislation will be considered as a comprehensive example, as it is rather strict. However, the discussed issues are essentially global in nature. The legal discussion here covers the main parts of the emerging regulatory framework for AI use by police in the EU: the General Data Protection Regulation (GDPR) [117] regulating the processing of personal data in most cases; the Law Enforcement Directive (LED) [118] regulating the processing of personal data for law enforcement purposes; the AI Act regulating the development and use of AI systems in high-risk contexts, though there are several exemptions for law enforcement purposes [119].

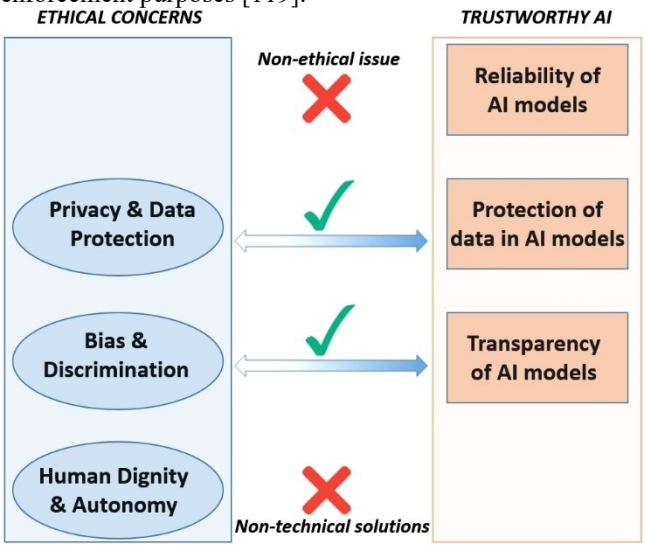

Fig. 6. A mapping of ethical concerns and technical Trustworthy AI means, regarding the large-scale deployment of AI by LEAs trying to combat illicit goods trafficking.

To provide a thematic discussion, ethical and privacy issues are discussed here alongside legal issues that represent the legislative manifestation of those ethical/privacy issues. There are three major areas where ethical and legal gray zones emerge, which are discussed below in Sections VI.A-VI.C. Potential technical countermeasures based on Trustworthy AI are subsequently presented in Sections VI.D-VI.G, while the mapping between ethical concerns and technical Trustworthy AI means is graphically summarized in Fig. 6.

A. Human Dignity and Autonomy

Human dignity entails that every human being possesses an inviolable intrinsic worth [120]; respecting this entails all people being treated as individual moral subjects, rather than mere objects [121]. Representing humans using AI reduces individuals to data points, thus affecting their dignity. Datafying people is a corporeal fetishistic interpretation of the body that abstracts individuals to an "arrangement of static code" [122]. This contributes to a real risk that operators begin to think of people as data points, which is a reductionist activity and can challenge the dignity of both operators and subjects of AI analysis.

Human dignity and autonomy could also be threatened by graph node classification, particularly from social network graphs and knowledge graphs. Node classification focuses on capturing social interactions to represent social relationships. Data produced through online activities, particularly social networks, are used to construct online "associative networks" [122]. Constructing digital "relationships" challenges the inherent complexity of human interactions, while simultaneously refusing to see an individual's autonomous associations. Associating to a community is often vital to constructing and expressing one's humanity and identity

[123], [124]. Instead, results of node classification, and community detection methodologies, can be viewed as representative and actionable, rather than the human-created autonomous reality.

From the legal perspective, automated decision-making that creates legal effects for data-subjects are prohibited unless an exception applies (GDPR, Article 22); though only adverse legal effects are prohibited in the law enforcement domain (LED, Article 11). Where, for example, investigators employ social network graphs during an investigation of illicit goods trafficking, they would need to assess the results of this data analysis carefully to ensure that they themselves are 'in the loop' of decision-making, rather than mistakenly delegating this to the AI system. As trafficking networks can be extensive, it is important that investigators do not forget that the masses of data they analyze using AI systems can represent individuals, and that their investigative decisions (or delegation of decisions) can have a real impact on those people.

B. Bias and Discrimination

In an AI context, equality demands that the system's operations cannot generate unfairly biased outputs [125]. AI systems should, ideally, not include historic bias. The continuation of such biases could lead to unintended (in)direct prejudice and discrimination [126]. Ensuring training and operational data is accurate, representative, reliable, relevant, and up-to-date with the population that the AI system will be applied to can be helpful in mitigating biases [127], while it is important for compliance with Article 10 of the EU's AI Act dealing with data governance. Yet, considering the massive, unsupervised gathering and analysis of both unverified and subjective data from the Surface Web, Dark Web, and social media sites when developing DNNs and graph analytics for identifying illicit goods trafficking, the potential for biased data to be included remains rife.

Analysis of social networks and knowledge graphs, particularly the Katz index, can result in "social sorting". This involves sorting individuals into categories, and assigning worth or risk to individuals [128]. Such profiling can constitute a serious data protection issue if the relevant safeguards are not followed [129]. This practice can also threaten fundamental rights such as dignity, equality, and integrity by classifying some individuals above or below others [128]. It is especially problematic where classifications are based on incorrect or biased information, resulting in discrimination [130]. This can then threaten the freedom of association and expression, and even incite behavioural change in the individuals' online actions, so as to avoid being profiled and sorted by such methodologies. Further, the use of AI-based profiling for law enforcement is a high-risk activity under the EU's AI Act Annex III, requiring greater scrutiny and regulation.

Predictive AI methods, like DNNs, can contribute to a feedback loop of biased outcomes. If DNNs are trained on biased data, there is a risk that historic biases continue or worsen with the technology deployment. This could have normative impacts, including, continuation and codification of discriminatory practices towards certain individuals, which could potentially exacerbate prejudice and marginalisation in society [126].

Consequently, we can see bias in AI systems as a socio-technical problem [131]. Therefore, solutions in the case of illicit goods trafficking need to take note of both social and technical issues and responses. Having interdisciplinary teams with an agreed and contextualized approach to fairness designing both the AI systems and comprehensive training materials with a culture that enabled open discussion about the identification and responses to various bias issues across the AI lifecycle can take significant steps to prevent discriminatory effects being created [131].

Where social network analysis and knowledge graphs are used in illicit trafficking investigations to visualize crawled data, the potentially very wide range of different data representing different people that might be processed means that it would be ideal for AI system providers to conduct bias reviews during deployment with social scientists who have expertise on bias/fairness in the contexts of deployment to ensure that biases have not been missed or misunderstood [131].

C. Privacy and Data Protection

The processing of crawled open-source data to train DNNs has the potential to infringe on people's privacy, engage research ethics considerations from an ethical perspective, and from a legal perspective engage both data-subject rights, and data protection regulations due to the vast amounts of publicly available personal data that can be gathered. Where information related to an identifiable person is subject to any operation, personal data is being processed (GDPR, Article 4(1) and 4(2)); this might include, for example, collecting details of cryptocurrency addresses/wallets that can be linked to their owners (even if this is very difficult to do).

The processing of personal data collected via crawling for DNN development and deployment can infringe upon people's privacy because it takes information from one context where people are happy to share it, or it is known to a small amount of people, and uses it in a different context where people might not want to accept this [132]. Depending on the data analyzed, training or using AI systems for combatting illicit trafficking could infringe upon the privacy of people's behaviour and action, their communications, their data and images. Depending on which data is gathered, the privacy of a person's thoughts and feelings (if shared), their location and space, and their associations could also be infringed upon [133]. Internet users might not reasonably expect their data to be collected and used to train AI systems. A research ethics committee might be reluctant to approve an activity that infringes upon people's privacy where they have a high expectation of privacy [134]. Private online spaces should be considered private insofar as a user has a reasonable expectation that they control who sees the information they share within that space; this is especially the case where site rules give the perception that users can tightly control access to their data. As such, the development of DNNs using crawled data needs to be done very carefully, with respect for

the people whose data is being processed, and a particular focus on transparency of the crawling for the internet users; this could involve publicising crawling efforts to make data-subjects aware of the processing. This would need to be re-planned or balanced against law enforcement priorities if, for example, the provision of transparency information to website users would alert them to police interest in their activities and this would frustrate an investigation by giving offenders an opportunity to destroy potentially incriminating evidence. Though this is less of a concern for LEA operations where covert infiltration of online groups can be essential to a successful investigation [135].

A key legal issue particularly in Europe, but also around the world [127] concerns identifying an appropriate legal basis for crawling open-source personal data to develop AI systems. The large number of potential data-subjects make it impractical to gather consent at scale for research ethics or legal basis purposes. However, the public task legal basis (GDPR, Article 6(1)(e)) might be allowed in some national laws. The legitimate interest legal basis could also be relevant where the processing is for scientific research, and does not override "the interests … of the data subject" (GDPR, Article 6(1)(f)); though compliance requires a legitimate interest assessment that very much depends on the reasonable expectations of the persons concerned, and can be very difficult to demonstrate [136]. Whatever legal basis is deemed appropriate, a data protection impact assessment (DPIA) is likely recommended in all cases of crawling due to processing personal data on a large scale, and is likely required where that data could be sensitive, as would be the case for data related to potential criminality [137]. However, the use of crawling during European LEA investigations is less legally problematic, as law enforcement purposes offer a very wide legal basis under the EU's LED, for example (LED, Article 1(1)). Yet, there could still be significant risks posed to data-subject through crawling their data and so a DPIA should still be considered for such purposes.

D. *Trustworthy and Technically Robust AI for Ethical Countermeasures*

The widespread deployment of the methods presented in Sections IV and V has the potential to bring about various detrimental side-effects for both citizens and organization. In order to proactively mitigate similar dangers, several initiatives have been set up to establish the principles of a trustworthy and secure AI. For example, the European Commission (EC) has created an independent expert group, namely the High-Level Expert Group on Artificial Intelligence, which prepared the "Ethics Guidelines for Trustworthy Artificial Intelligence (AI)" document [138]. It is a framework that aims to secure fundamental citizen rights, formulate ethical principles and make current AI systems trustworthy from societal, ethical, and legal perspectives. The same year, ACM released the ACM Code of Ethics and Professional Conduct [139], covering ethics across the computing field: the tech code of ethics, computing ethics, software ethics, programming ethics, AI ethics, etc. Prior to this, IEEE launched the IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems with the mandate to provide a practical guide [140] for addressing the urgent relevant ethical concerns. All these initiatives paved the way for many ethical frameworks [141] regarding trustworthy AI in various domains, such as digital healthcare [142], telecommunications (6G) [143], or fintech [144], as well as for many types of sectors, e.g., public, private and non-governmental organizations [145]. In order to provide technically robust and trustworthy AI systems, potential solutions should focus on three key aspects, which are detailed below.

E. *Transparency of AI Models*

Transparency can be achieved through distinct interpretability and explainability features, which should be employed for both the design of the AI models and the description of the data used during model conceptualization. Transparency allows software developers to analyze their system, enables security practitioners to trust it and facilitates regulators to ensure it is safe and fair. Unfortunately, in many cases, these AI models become so complex that it is often challenging to understand the reasons behind their predictions/decisions. In general, "black-box" models yield excellent performance in terms of accuracy but users, their designers, cannot fully understand how the underlying variables are combined to reach a decision. This fact brings to the fore the risk of not respecting the fundamental rights of citizens to privacy preservation, to equality before the law, etc. Thus, algorithms have been developed for interpreting and clarifying the decision-making process of AI models, creating the new subfield of explainable AI (XAI) [146]. In this context, new regulations have sprung up requiring human-readable reports of how AI systems deployed in the real world have arrived at their conclusions. One notable example is Europe's GDPR, which stipulates that organizations must disclose "meaningful information about the logic involved" when using personal data for "automated decision-making, including profiling". All these laws have been motivated by the constantly growing concern that "black-box" systems may be hiding evidence of illegal, or perhaps just unsavory, discriminatory practices [147]. Thus, it is evident how making XAI deployment mandatory by law is pertinent for fostering socially responsible LEA practices.

In order to cope with the aforementioned transparency problems and to comply with EU regulations, engineers are encouraged to follow an explainability-by-design approach during the system's architectural design [148]. Moreover, development and use of inherently explainable "glass-box" AI models (e.g., models derived from causal analysis, or relying on graph reasoning) should also be encouraged over "black-box" models (e.g., DNNs), in cases where the former ones yield acceptable performance [149]. In cases where "black-box" models are deemed necessary (e.g., due to their high accuracy when large data volumes are available), their explainability must be targeted via well-established XAI algorithms, able to interpret model predictions and boost trustworthiness. Examples of such algorithms include Deep Learning Important Features – DeepLift [150], Local

Interpretable Model-agnostic Explanations – LIME [151], Shapley Additive Explanations – SHAP [152] and Causal Explanations – CXPlain [153].

XAI can be leveraged to explain to security practitioners the reasoning behind a predicted mapping from illicit goods trafficking activities to specific criminal networks and, thus, guarantee lawful prosecution through court-proof evidence collected by LEAs. The overall ambition of this process is to remove concerns about false matches, low fairness and reliability caused by bias mitigation in the utilized AI algorithms. For instance, XAI can provide transparency in how AI identifies suspicious transactions involving precursor chemicals, ensuring that decisions can be audited and justified. A main concern is to exploit this transparency for identifying and avoiding the perpetuation of systemic social inequalities, such as racial and gender discrimination, marginalization, etc., which may be reproduced by AI model predictions due to the data they have been trained on. An example would be XAI methods that directly let humans know whether an AI prediction was made due to inherent model bias (e.g., identifying a person as potential criminal based on skin color). This functionality can be utilized even during the initial construction of AI models, to ensure that the training data are diverse, representative, and free from bias against minorities. Notably, implicitly fusing multiple different information sources on a Knowledge Graph (as discussed in Section V.C) has the potential to automatically mitigate such biases that are present only in a subset of the utilized input modalities. However, explicitly enforcing AI transparency at the previous stages of the analysis significantly enhances overall trustworthiness and facilitates better compliance with regulations and laws.

XAI can be offered as a library of methods aiming to explain how "glass-box" recommendations are made, or even "black-box" ones up to a degree. However, in complex problems such as the accurate identification of illicit goods trafficking networks, the trade-offs between accuracy and interpretability must be carefully balanced: explaining "black-box" AI model predictions is particularly challenging and non-trivial from a technical standpoint.

F. Reliability of AI Models

AI reliability involves identifying in early stage any potential vulnerabilities and implementing appropriate technical solutions to guarantee that the overall system will not fail or be manipulated by an adversary [154]. In general, poor real-world accuracy of AI systems and the identification of vulnerabilities leading to malfunctions are the key indicators of AI model unreliability. In AI-enabled systems for fighting illicit goods trafficking, the main risk lies in too high rates of false positives and/or false negatives, when detecting illicit items/activities, as well as in erroneous correlations between such activities and criminal networks and/or individuals. Overcoming this kind of issues is a prerequisite for AI trustworthiness, but it is not trivial.

In machine learning systems, such as DNNs, overfitting is a major challenge to reliability during real-world deployment, due to low generalization ability in novel data inputs. The issue arises when the AI model has memorized in detail the noise or the random error components within its training dataset, instead of learning the actually desired patterns. Dropout methods [155] and data augmentation [156] algorithms can pre-emptively mitigate such problems during the training phase. Dropout essentially turns the trained DNN into an ensemble of multiple subnetworks with different topologies, thus combatting overfitting by averaging their predictions. Data augmentation simply involves artificially increasing the size of the training dataset, while batch normalization standardizes the input of each DNN layer during training to zero mean and unit variance, resulting in a regularizing effect due to the insertion of noise. All these approaches are applied after carefully reducing the complexity of the AI model by eliminating layers or neurons, to minimize susceptibility to overfitting. A complementary aspect of DNN reliability during deployment, also related to maintaining high generalization ability in novel data inputs, is achieving robustness to noisy inputs. A particular, "extreme" instance of this problem is robustness against adversarial attacks, i.e., a particular scenario where a pretrained DNN is fed special inputs with minimal noise, carefully and explicitly added so that the AI model is fooled. Currently, properly adapting the DNN training objectives seems to be the best pre-emptive strategy for ensuring deployment-time robustness [157].

Another aspect of AI reliability is the elimination of any potential bias in the algorithms, which can be detected using XAI approaches. Notably, the terms bias, discrimination and unfairness are often used interchangeably with similar meanings. AI bias is becoming more apparent and problematic with the wider use of AI-based decision support systems. One of its negative consequences is discrimination, which refers to the unfair or unequal treatment of individuals based on certain characteristics [158]. Special algorithmic mechanisms can be utilized to ensure the fairness of AI models, with the simplest ones being systematic methods to curate the training datasets for avoiding biases that may unfairly target specific demographics or regions. Technical schemes such as data augmentation and data balancing can be exploited to ensure fair representation from different categories of illicit goods and trafficking scenarios. The relevant literature often links discrimination with bias, which refers to a deviation from the standard that is necessary for identifying statistical patterns in the data. Simultaneously, reducing bias and discrimination is directly linked to the perception of justice [159]. Understanding AI decision-making processes becomes paramount for justifying decisions, especially in safety and security cases. However, these processes often do not provide information relevant to judgments of justice, which can compromise the fundamental accountability of human decision-making.

In the context of the methods presented in Sections IV and V, AI can be used in courts to assess whether a defendant has committed or will recommit a crime, in criminal identification through facial recognition from CCTV, etc. [160]. Given that such applications of AI have direct consequences in citizens' lives and can become harmful if designed without taking into considerations the fairness principle [161], [162], the value of AI reliability becomes evident. Yet, several studies have

shown that satisfying multiple notions of fairness simultaneously is an impossible task; individual and group fairness may occasionally be incompatible [163], [164].

From a legal perspective, reliability is crucial to criminal trials as AI systems that analyse data in a reliable and reproducible way are essential to ensuring a fair trial [165]. Yet, under the EU's AI Act, reliability is not specifically covered. It could be considered as part of robustness under Article 15. However, in line with current digital forensic practices where experts provide their opinion on the reliability of the evidence [165], it could also be considered under human oversight though Article 14. In any case, the absence of specific consideration of reliability in the EU's AI Act does not affect its importance in other legal areas, and the need of AI developers to respect this, especially as fair trial considerations are of paramount importance where AI systems are used in the criminal justice system.

G. *Protection of Data in AI Models*

Data protection is achievable through applying appropriate control measures in order to ensure the citizens' right to privacy. It is essential to address the possibility of deployed AI models having unintentionally memorized personal details unrelated to their primary goal, which could lead to unintended disclosure [166]. This can be problematic, especially when the AI models have been trained with private or sensitive datasets. Sophisticated methods that guarantee data privacy have emerged in the context of trustworthy AI and can be readily applied. For instance, in the EU, anonymization or pseudonymization methods are used to support GDPR compliance of AI systems. Advanced AI algorithms, e.g., for face de-identification in image datasets [167], can be exploited for enhancing privacy guarantees, while differential privacy methods may be deployed to prevent re-identification of individuals through the AI system's outputs [168]. Moreover, privacy-by-design AI methods such as federated DNN learning can be employed to protect sensitive data [169], while allowing collaboration between different LEAs. The need for data protection increases further when additional modalities and information sources are utilized to improve functionality (e.g., [170]). In general, technical solutions (e.g., anonymizing data during Surface Web crawling and social media analysis, using encryption and secure storage for Dark Web and network traffic analysis, applying differential privacy in cryptocurrency transaction analysis and LEA database processing, etc.) can prevent unintended disclosure of personal information when deploying AI systems for combatting illicit goods trafficking. However, besides technical measures, organizations can complete a DPIA to evaluate privacy and data protection impacts of AI systems, even during the system design phase, as mentioned above. The goal is to identify and address potential issues associated with high-risk instances of data processing; in this case, with a particular emphasis on applications related to public security and use by LEAs.

As it can be seen in Fig. 6, ethical concerns and technical Trustworthy AI means overlap but do not map to each other completely. Issues of human dignity and autonomy are more readily dealt with via a suitable legal, cultural and operational framework. On the other hand, the reliability of AI models that are deployed in the field by LEAs, although legally relevant, is primarily a technical (not ethical) concern. Finally, technical means for transparent AI and for data protection in AI are able to directly address certain relevant ethical concerns. Therefore, it is necessary to combine Trustworthy AI with an appropriate legal/cultural/operational environment, to successfully deploy AI-enabled solutions with the potential for large-scale social and operational acceptance.

VII. Conclusion

Recent high-tech trends employed by organized crime for illicit goods trafficking can be effectively countered using advanced AI methods tailored for large-scale information processing. This is critical due to the growing digital sophistication of trafficking networks and the significant social, economic, and environmental harm they cause. The deployment of such AI tools by LEAs is a practical means to address the emerging criminal modi operandi identified in this paper. While various legal and ethical considerations inevitably arise, a subset of these issues can be mitigated through technical means, particularly within the Trustworthy AI framework. These means should be seamlessly integrated into a broader set of regulations and procedures to ensure ethical, legal, and effective AI-powered law enforcement.

This paper contributes to the existing literature by offering a novel interdisciplinary perspective that integrates criminology, AI, and legal/ethical studies into a comprehensive framework. It extends current understanding by suggesting practical mitigation strategies, such as the application of Explainable AI (XAI) for transparency, privacy-preserving methods to protect sensitive data, and federated learning for secure collaboration among LEAs. Additionally, it advocates for the further development and exploration of knowledge graphs for fusing diverse information sources, enhancing AI applications in law enforcement. By emphasizing the importance of fairness, the need for advanced bias mitigation techniques, and the development of a robust regulatory framework, the paper lays the groundwork for future research in developing ethical and reliable AI systems to combat illicit goods trafficking.

By addressing these challenges through the lens of Trustworthy AI, this study also highlights the critical need for advanced AI methods in responding to the rapidly evolving strategies of illicit goods trafficking. For governments, LEAs, and commercial entities, these tools are not just theoretical; they have immediate, practical applications. For instance, AI-driven cryptocurrency transaction analysis can unmask the financial networks underpinning human trafficking, while social network graph analysis aids in dismantling organized crime by identifying key actors and hidden links. The full adoption of Trustworthy AI ensures transparency and privacy preservation, while building public trust and safeguarding citizens' rights. These innovations empower LEAs to respond swiftly and effectively to crimes that harm global economies and societies, underscoring the urgency of implementing such solutions worldwide.

Future work should focus on further refining these strategies and exploring how AI can be integrated into LEA practices in

a way that is both effective and compliant with ethical and legal standards. As technology and criminal methods continue to evolve, it is crucial to advance AI methodologies that can adapt to new challenges while maintaining a focus on protecting citizens' rights and ensuring societal trust.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101073876 (Ceasefire). This publication reflects only the authors' views. The European Commission is not responsible for any use that may be made of the information it contains.

REFERENCES

[1] Europol, European Union, "Serious and organized crime threat assessment, a corrupting influence: the infiltration and undermining of Europe's economy and society by organized crime", Publications Office of the European Union, Luxembourg, 2021.

[2] World Population Review. (2023). *Gun Deaths by Country*. [Online]. Available: https://worldpopulationreview.com/country-rankings/gun-deaths-by-country.

[3] European Monitoring Centre for Drugs and Drug Addiction and Europol, "Drug-related deaths and mortality in Europe: update from the EMCDDA expert network", Publications Office of the European Union, Luxembourg, 2021.

[4] Europol. (2022). *Business Fundamentals: How Illegal Drugs Sustain Organised Crime in Europe*. [Online]. Available: https://www.europol.europa.eu/cms/sites/default/files/documents/business_fundamentals_how_illegal_drugs_sustain_organised_crime_in_europe.pdf

[5] European Monitoring Centre for Drugs and Drug Addiction and Europol, "EU Drug Markets Report 2019", Publications Office of the European Union, Luxembourg, 2019.

[6] D. Petkovic, "It is not 'Accuracy vs. Explainability'—we need both for trustworthy AI systems," *IEEE Trans. Technol. Soc.*, vol. 4, no. 1, pp. 46–53, 2023, doi: 10.1109/TTS.2023.3239921.

[7] J.R. Schoenherr, R. Abbas, K. Michael, P. Rivas, T.D. Anderson, "Designing AI using a human-centered approach: Explainability and accuracy toward trustworthiness," *IEEE Trans. Technol. Soc.*, vol. 4, no. 1, pp. 9–23, 2023, doi: 10.1109/TTS.2023.3257627.

[8] D. Abate, M. Paolanti, and R. Pierdicca, A. Lampropoulos, K. Toumbas, A. Agapiou, S. Vergis, E. Malinverni, K. Petrides, A. Felicetti, et al., "SIGNIFICANCE: Stop illicit heritage trafficking with Artificial Intelligence", *ISPRS Congress,* Nice, France, 2022, pp. 729–736.

[9] M. Orantes, "Leveraging machine learning and Artificial Intelligence to combat human trafficking", dissertation, Utica College, Utica, NY, USA, 2018.

[10] M. Hernández-Álvarez, "Detection of Possible Human Trafficking in Twitter", in *Proc. ICI2ST*, Quito, Ecuador, 2019, pp. 187-191.

[11] M. B. Sarwar, M. K. Hanif, R. Talib, M. Younas and M. U. Sarwar, "DarkDetect: Darknet Traffic Detection and Categorization Using Modified Convolution-Long Short-Term Memory", *IEEE Access*, vol. 9, pp. 113705–113713, 2021, doi: 10.1109/ACCESS.2021.3105000.

[12] K. J. Hayward and M. M. Maas, "Artificial Intelligence and crime: A primer for criminologists", *Crime Media Cult.*, vol. 17, no. 2, pp. 209–233, 2021, doi: 10.1177/1741659020917434.

[13] J. Deeb-Swihart, A. Endert and A. Bruckman, "Ethical Tensions in Applications of AI for Addressing Human Trafficking: A Human Rights Perspective," *Proc. ACM Hum.-Comput. Interact.,* vol. 6, no. CSW2, pp. 1–29, 2022, doi: 10.1145/3555186.

[14] T. Rademacher, "Artificial Intelligence and law enforcement," in *Regulating Artificial Intelligence*, Springer, Cham, 2020, pp. 225–254.

[15] G. R. Newman, "Cybercrime," in *Handbook on Crime and Deviance*, M. D. Krohn, A. J. Lizotte, G. Penly Hall, Eds. New York, NY, USA: Springer, 2009, pp. 551–584.

[16] J. Aldridge, "Does online anonymity boost illegal market trading?" *Media Cult. Soc.*, vol. 41, pp. 578–583, 2019, doi: 10.1177/0163443719842075.

[17] H. Chen, *Dark Web: Exploring and Data Mining the Dark Side of the Web*. New York, NY, USA: Springer-Verlag, 2012.

[18] J. Martin, "Lost on the Silk Road: Online drug distribution and the 'cryptomarket'," *Criminol. Crim. Justice*, vol. 14, pp. 351–367, 2014, doi: 10.1177/1748895813505234.

[19] G. P. Persi, J. Aldridge, R. Nathan, R. Warnes, "Behind the curtain: The illicit trade of firearms, explosives and ammunition on the dark web," RAND Corporation, 2017. [Online]. Available: https://www.rand.org/pubs/research_reports/RR2091.html

[20] K. Kruithof, J. Aldridge, D. D. Hétu, M. Sim, E. Dujso, S. Hoorens, "Internet-facilitated drugs trade: An analysis of the size, scope and the role of the Netherlands," RAND Corporation, 2016. [Online]. Available: https://www.rand.org/pubs/research_reports/RR1607.html

[21] Europol, European Union, "Internet organised crime threat assessment 2021," Publications Office of the European Union, Luxembourg, 2021.

[22] UNODC, "COVID-19 and the drug supply chain: From production and trafficking to use," United Nations Office on Drugs and Crime, Vienna, 2020.

[23] J. Aldridge, R. Askew, "Delivery dilemmas: How drug cryptomarket users identify and seek to reduce their risk of detection by law enforcement," *Int. J. Drug Policy*, vol. 41, pp. 101–109, 2017, doi: 10.1016/j.drugpo.2016.10.010.

[24] N. Christin, "An EU-focused analysis of drug supply on the AlphaBay marketplace," EMCDDA, 2017.

[25] N. Christin, "Traveling the Silk Road: A measurement analysis of a large anonymous online marketplace," in *Proc. WWW*, Rio de Janeiro, Brazil, pp. 213–224, 2013.

[26] D. Décary-Hétu, O. Quessy-Doré, "Are repeat buyers in cryptomarkets loyal customers? Repeat business between dyads of cryptomarket vendors and users," *Am. Behav. Sci.*, vol. 61, no. 11, pp. 1341–1357, 2017, doi: 10.1177/0002764217734265.

[27] J. Aldridge, D. Décary-Hétu, "Not an 'Ebay for drugs': The cryptomarket 'Silk Road' as a paradigm shifting criminal innovation," *SSRN Electron. J.*, 2014, doi: 10.2139/ssrn.2436643.

[28] MM. Gilbert, N. Dasgupta, "Silicon to syringe: Cryptomarkets and disruptive innovation in opioid supply chains," *Int. J. Drug Policy*, vol. 46, pp. 160–167, 2017, doi: 10.1016/j.drugpo.2017.05.052.

[29] Frontex, "Risk analysis for 2022/2023," Warsaw, Poland, 2022. [Online]. Available: https://prd.frontex.europa.eu/document/risk-analysis-for-2022-2023/.

[30] UNODC, "The illicit market in firearms," United Nations Office on Drugs and Crime, Vienna, 2019.

[31] D. Décary-Hétu, M. Paquet-Clouston, J. Aldridge, "Going international? Risk taking by cryptomarket drug vendors," *Int. J. Drug Policy*, vol. 35, pp. 69–76, 2016, doi: 10.1016/j.drugpo.2016.06.003.

[32] M. Tzanetakis, G. Kamphausen, B. Werse, R. von Laufenberg, "The transparency paradox: Building trust, resolving disputes and optimising logistics on conventional and online drugs markets," *Int. J. Drug Policy*, vol. 35, pp. 58–68, 2016, doi: 10.1016/j.drugpo.2015.12.010.

[33] European Monitoring Centre for Drugs and Drug Addiction and Europol, "Drugs and the darknet: Perspectives for enforcement, research and policy," Publications Office of the European Union, Luxembourg, 2017.

[34] Europol, "European Union serious and organised crime threat assessment: A corrupting influence: The infiltration and undermining of Europe's economy and society by organised crime," Publications Office of the European Union, Luxembourg, 202.

[35] N. Miotto, "The role of online communities in supporting 3D-printed firearms," GNET (blog), August 25, 2021. Available: https://gnet-research.org/2021/08/25/the-role-of-online-communities-in-supporting-3d-printed-firearms/.

[36] M. Dressler and N. Duquet, "Illicit firearms trafficking in Europe during and after COVID-19," News Article, Flemish Peace Institute, 2020. Available: https://vlaamsvredesinstituut.eu/en/newspost/illicit-firearms-trafficking-in-europe-during-and-after-covid-19/

[37] European Monitoring Centre for Drugs and Drug Addiction and Europol, "The Internet and drug markets," Publications Office of the European Union, Luxembourg, 2016.

[38] M. J. Barratt, J. A. Ferris, A. R. Winstock, "Use of Silk Road, the online drug marketplace, in the United Kingdom, Australia and the United States," *Addiction*, vol. 109, pp. 774–783, 2014, doi: 10.1111/add.12470.

[39] E. Lahaie, M. Martinez, A. Cadet-Taïrou, "New psychoactive substances and the internet: Current situations and issues," *Tendances*, vol. 84, Observatoire Français des Drogues et des Toxicomanies (OFDT), 2013.

[40] European Monitoring Centre for Drugs and Drug Addiction and Europol, "EMCDDA special report: COVID-19 and drugs—Drug supply via darknet markets," Publications Office of the European Union, Luxembourg, 2020.

[41] A. Bergeron, D. Décary-Hétu, L. Giommoni, M. P. Villeneuve-Dubuc, "The success rate of online illicit drug transactions during a global pandemic," *Int. J. Drug Policy*, vol. 99, 2022, doi: 10.1016/j.drugpo.2021.103452.

[42] M. Chawki, "The dark web and the future of illicit drug markets," *J. Transp. Secur.*, vol. 15, pp. 173–191, 2022, doi: 10.1007/s12198-022-00252-y.

[43] R. Rawat, A. S. Rajawat, V. Mahor, R. N. Shaw, and A. Ghosh, "Dark Web: onion hidden service discovery and crawling for profiling morphing, unstructured crime and vulnerabilities prediction," in *Proc. ICEEE*, pp. 717–734, 2021.

[44] I. N. Rezende, "Facial recognition in police hands: Assessing the 'Clearview case' from a European perspective", *New J. Eur. Crim. Law*, vol. 11, no. 3, pp. 375–389, 2020, doi: 10.1177/2032284420948161.

[45] G. Batsis, I. Mademlis, and G. T. Papadopoulos, "Illicit item detection in X-ray images for security applications," in *Proc. BDS*, Athens, Greece, 2023, pp. 63–70.

[46] P. Yadav, N. Gupta, and P. K. Sharma, "A comprehensive study towards high-level approaches for weapon detection using classical machine learning and deep learning methods," *Expert Syst. Appl.*, vol. 212, pp. 118698, 2023, doi: 10.1016/j.eswa.2022.118698.

[47] N. Dwivedi, D. K. Singh and D. S. Kushwaha, "Weapon Classification using Deep Convolutional Neural Network," in *Proc. IEEE ICICT*, Allahabad, India, 2019, pp. 1–5.

[48] V. Kaya, S. Tuncer, A. Baran, "Detection and classification of different weapon types using deep learning," *Appl. Sci.*, vol. 11, no. 16, pp. 7535, 2021, doi: 10.3390/app11167535.

[49] A. Warsi, M. Abdullah, M. N. Husen, M. Yahya, S. Khan, and N. Jawaid, "Gun Detection System Using YOLOv3," in *Proc. IEEE ICSIMA*, Kuala Lumpur, Malaysia, 2019, pp. 1–4.

[50] A. Warsi, M. Abdullah, M. N. Husen, M. Yahya, "Automatic Handgun and Knife Detection Algorithms: A Review", in *Proc. IMCOM*, Taichung, Taiwan, 2020, pp. 1–9.

[51] H. Jain, A. Vikram, Mohana, A. Kashyap and A. Jain, "Weapon Detection using Artificial Intelligence and Deep Learning for Security Applications," in *Proc. ICESC*, Coimbatore, India, 2020, pp. 193–198.

[52] M. T. Bhatti, M. G. Khan, M. Aslam and M. J. Fiaz, "Weapon Detection in Real-Time CCTV Videos Using Deep Learning," *IEEE Access*, vol. 9, pp. 34366–34382, 2021, doi: 10.1109/ACCESS.2021.3059170.

[53] D. Romero and C. Salamea, "Convolutional models for the detection of firearms in surveillance videos," *Appl. Sci.*, vol. 9, no. 15, pp. 2965, doi: 10.3390/app9152965.

[54] J. Salido, V. Lomas, J. Ruiz-Santaquiteria, and O. Deniz, "Automatic handgun detection with deep learning in video surveillance images," *Appl. Sci.*, vol. 11, no. 13, pp. 6085, 2021, doi: 10.3390/app11136085.

[55] A. Lamas, S. Tabik, A. C. Montes, F. Pérez-Hernández, J. García, R. Olmos, and F. Herrera, "Human pose estimation for mitigating false negatives in weapon detection in video-surveillance," *Neurocomputing*, vol. 489, 2022, pp. 488–503, doi: 10.1016/j.neucom.2021.12.059.

[56] A. Egiazarov, V. Mavroeidis, F. M. Zennaro and K. Vishi, "Firearm Detection and Segmentation Using an Ensemble of Semantic Neural Networks," in *Proc. EISIC*, Oulu, Finland, 2019, pp. 70–77.

[57] L. Pang, H. Liu, Y. Chen, and J. Miao, "Real-time concealed object detection from passive millimeter-wave images based on the YOLOv3 algorithm", *Sensors*, vol. 20, pp. 1678, 2020, doi: 10.3390/s20061678.

[58] W. Zhang, Q. Zhu, Y. Li, and H. Li, "MAM Faster R-CNN: Improved Faster R-CNN based on Malformed Attention Module for object detection on X-ray security inspection," *Digit. Signal Process.*, vol. 2, pp. 104072, 2023, doi: 10.1016/j.dsp.2023.104072.

[59] Y. Wei, R. Tao, Z. Wu, Y. Ma, L. Zhang, and X. Liu, "Occluded prohibited items detection: An X-ray security inspection benchmark and de-occlusion attention module," In *Proc. ACM MM*, Seattle, USA, 2020, pp. 138–146.

[60] R. Tao, Y. Wei, X. Jiang, H. Li, H. Qin, J. Wang, Y. Ma, L. Zhang, and X. Liu, "Towards real-world X-ray security inspection: A high-quality benchmark and lateral inhibition module for prohibited items detection," in *Proc. IEEE/CVF ICCV*, Montreal, QC, Canada, 2021, pp. 10903–10912.

[61] C. Miao, L. Xie, F. Wan, C. Su, H. Liu, J. Jiao, and Q. Ye, "SIXray: A Large-Scale Security Inspection X-Ray Benchmark for Prohibited Item Discovery in Overlapping Images," *Proc. IEEE/CVF CVPR*, Long Beach, CA, USA, 2019, pp. 2114–2123.

[62] B. Ma, T. Jia, M. Su, X. Jia, D. Chen, and Y. Zhang, "Automated Segmentation of Prohibited Items in X-ray Baggage Images Using Dense De-overlap Attention Snake," *IEEE Trans. Multimedia*, vol. 25, pp. 4374–4386, 2022, doi: 10.1109/TMM.2022.3174339.

[63] X. Wang, P. Peng, C. Wang, and G. Wang, "You are your photographs: Detecting multiple identities of vendors in the darknet marketplaces," In *Proc. ASIACCS*, Incheon, Korea, 2018, pp. 431–442.

[64] J. Li, Q. Xu, N. Shah, T.K. Mackey, "A machine learning approach for the detection and characterization of illicit drug dealers on Instagram: model evaluation study," *J. Med. Internet Res.*, vol. 21, no. 6, pp. e13803, 2019, doi: 10.2196/13803.

[65] R. Li, M. Tobey, M. E. Mayorga, S. Caltagirone, and O. Y. Özaltın, "Detecting Human Trafficking: Automated Classification of Online Customer Reviews of Massage Businesses," *Manuf. Serv. Oper. Manag.*, vol. 25, no. 3, pp. 1051–1065, 2023, doi: 10.1287/msom.2023.1196.

[66] C. Hu, B. Liu, Y. Ye, and X. Li, "Fine-grained classification of drug trafficking based on Instagram hashtags", *Decis. Support Syst.*, vol. 165, pp. 113896, 2023, doi: 10.1016/j.dss.2022.113896.

[67] C. Hu, M. Yin, B. Liu, X. Li, and Y. Ye, "Identifying Illicit Drug Dealers on Instagram with Large-Scale Multimodal Data Fusion," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 5, pp. 1–23, 2021, doi:10.1145/3472713.

[68] M. W. Al-Nabki, E. Fidalgo, E. Alegre, and L. Fernández-Robles, "Improving named entity recognition in noisy user-generated text with local distance neighbor feature," *Neurocomputing*, vol. 382, pp. 1–11, 2020, doi: 10.1016/j.neucom.2019.11.072.

[69] B. Nie, C. Li, and H. Wang, "KA-NER: Knowledge-augmented Named Entity Recognition", In *Proc. CCKS*, Guangzhou, China, 2021, pp. 60–75.

[70] M. Pikuliak, M. Simko, and M. Bielikova, "Cross-lingual learning for text processing: A survey", *Expert Syst. Appl.*, vol. 165, pp. 113765, 2021, doi: 10.1016/j.eswa.2020.113765.

[71] M. B. Sarwar, M. K. Hanif, R. Talib, M. Younas and M. U. Sarwar, "DarkDetect: Darknet Traffic Detection and Categorization Using Modified Convolution-Long Short-Term Memory", *IEEE Access*, vol. 9, pp. 113705-113713, 2021, doi: 10.1109/ACCESS.2021.3105000.M.

[72] J. Lan, X. Liu, B. Li, Y. Li, T. Geng, "DarknetSec: A novel self-attentive deep learning method for darknet traffic classification and application identification," *Comput. Secur.*, vol. 116, pp. 102663, 2022, doi: 10.1016/j.cose.2022.102663.

[73] A. Mezina, R. Burget, and A. Ometov, "Reinterpreting Usability of Semantic Segmentation Approach for Darknet Traffic Analysis", *Comput. Netw.*, vol. 249, pp. 110493, 2024, doi: 10.1016/j.comnet.2024.110493.

[74] S. Al-E'mari, Y. Sanjalawe, S. Fraihat, "Detection of obfuscated Tor traffic based on bidirectional generative adversarial networks and vision transform," *Comput. Secur.*, vol. 135, pp. 103512, 2023, doi: 10.1016/j.cose.2023.103512.

[75] Y. Liu, X. Wang, B. Qu, and F. Zhao, "ATVITSC: A novel encrypted traffic classification method based on deep learning", *IEEE Trans. Inf. Forensics Secur.*, 2024, doi: 10.1109/TIFS.2024.3433446.

[76] A. Berman and C. L. Paul, "Making sense of Darknet markets: Automatic inference of semantic classifications from unconventional multimedia datasets," in *Proc. HCII*, Orlando, FL, USA, 2019, pp. 230–248.

[77] L. Choshen, D. Eldad, D. Hershcovich, E. Sulem and O. Abend, "The language of legal and illegal activity on the Darknet," *arXiv:1905.05543*, 2019.

[78] Y. Zhu, J. Tao, H. Wang, L.X. Yu, Y. Luo, T. Qi, Z. Wang, and Y. Xu, "DGNN: Accurate Darknet Application Classification Adopting Attention Graph Neural Network," *IEEE Trans. Netw. Serv. Manag.*, vol. 21, pp. 1660–1671, 2023, doi: 10.1109/TNSM.2023.3344580.

[79] J. Saleem, R. Islam, and M. Z. Islam, "Darknet traffic analysis: a systematic literature review," *IEEE Access*, vol. 12, pp. 42423–42452, 2024, doi: 10.1109/ACCESS.2024.3373769.

[80] P. J. Carrington, "Crime and Social Network Analysis," In *The SAGE Handbook of Social Network Analysis*, Los Angeles, CA, USA: SAGE Publications, 2011, pp. 236–255.

[81] K. von Lampe, *Organized Crime: Analyzing illegal activities, criminal structures, and extra-legal governance*, 1st ed. Thousand Oaks, CA, USA: SAGE Publications, 2016.

[82] D. Bright and C. Whelan, *Organised Crime and Law Enforcement: A Network Perspective*. London, UK: Routledge, 2020.

[83] Y. Qian, Y. Zhang, Y. Ye, and C. Zhang, "Distilling meta knowledge on heterogeneous graph for illicit drug trafficker detection on social media," in *Proc. NIPS*, Virtual Conference, 2021, pp. 26911–26923.

[84] P. A. C. Duijn and P. P. H. M. Klerks, "Social network analysis applied to criminal networks: Recent developments in Dutch law enforcement," in

Networks and Network Analysis for Defence and Security, Lecture Notes in Social Networks. Cham, Switzerland: Springer, 2014, pp. 121–159.
[85] S.-U. Hassan, M. Shabbir, S. Iqbal, A. Said, F. Kamiran, R. Nawaz, and U. Saif, "Leveraging deep learning and SNA approaches for smart city policing in the developing world," *Int. J. Inf. Manage.*, vol. 56, p. 102045, 2021, doi: 10.1016/j.ijinfomgt.2019.102045.
[86] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, "Fast unfolding of communities in large networks", *J. Stat. Mech.*, p. 10008, 2008, doi: 10.1088/1742-5468/2008/10/P10008.
[87] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proc. Natl. Acad. Sci. USA*, vol. 99, no. 12, pp. 7821–7826, 2002, doi: 10.1073/pnas.122653799.
[88] A. Tundis, A. Jain, G. Bhatia, and M. Mühlhäuser, "Similarity analysis of criminals on social networks: An example on Twitter," in *Proc. ICCCN*, Valencia, Spain, 2019, pp. 1–9.
[89] M. W. Al-Nabki, E. Fidalgo, E. Alegre, and L. Fernández-Robles, "ToRank: Identifying the most influential suspicious domains in the Tor network," *Expert Syst. Appl.*, vol. 123, pp. 212–226, 2019, doi: 10.1016/j.eswa.2019.01.029.
[90] A. Bahulkar, B. K. Szymanski, N. O. Baycik, and T. C. Sharkey, "Community detection with edge augmentation in criminal networks," in *Proc. ASONAM*, Barcelona, Spain, 2018, pp. 1168–1175.
[91] F. Calderoni, S. Catanese, P. De Meo, A. Ficara, and G. Fiumara, "Robust link prediction in criminal networks: A case study of the Sicilian Mafia," *Expert Syst. Appl.*, vol. 161, p. 113666, 2020, doi: 10.1016/j.eswa.2020.113666.
[92] W. Lin, S. Ji, and B. Li, "Adversarial attacks on link prediction algorithms based on graph neural networks," in *Proc. ASIA CCS*, Taipei, Taiwan, 2020, pp. 370–380.
[93] M. Lim, A. Abdullah, N. Jhanjhi, and M. Khurram Khan, "Situation-aware deep reinforcement learning link prediction model for evolving criminal networks," *IEEE Access*, vol. 8, pp. 16550–16559, 2020, doi: 10.1109/ACCESS.2019.2961805.
[94] R. V. Gundur, M. Levi, V. Topalli, M. Ouellet, M. Stolyarova, L. Y.-C. Chang, and D. D. Mejía, "Evaluating criminal transactional methods in cyberspace as understood in an international context," *CrimRxiv*, Apr. 2021, doi: 10.21428/cb6ab371.5f335e6f.
[95] M. Ahmed, I. Shumailov, and R. Anderson, "Tendrils of crime: Visualizing the diffusion of stolen bitcoins," in *Proc. GraMSec*, Oxford, UK, 2018, pp. 1–16.
[96] L. Serena, S. Ferretti, G. D'Angelo, "Cryptocurrencies activity as a complex network: Analysis of transactions graphs," *Peer-to-Peer Netw. Appl.*, vol. 15, pp. 839–853, 2022, doi: 10.1007/s12083-021-01220-4.
[97] H. Tian, Y. Li, Y. Cai, X. Shi, and Z. Zheng, "Attention-based graph neural network for identifying illicit Bitcoin addresses," in *Proc. BlockSys*, Guangzhou, China, 2021, pp. 147–162.
[98] J. Liu, J. Zheng, J. Wu, Z. Zheng, "FA-GNN: Filter and augment graph neural networks for account classification in Ethereum," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 4, pp. 2579–2588, 2022, doi: 10.1109/TNSE.2022.3166655.
[99] S. Meiklejohn, M. Pomarole, G. Jordan, K. Levchenko, D. McCoy, G. M. Voelker, and S. Savage, "A fistful of Bitcoins: Characterizing payments among men with no names," *Commun. ACM*, vol. 59, no. 4, pp. 86–93, 2016, doi: 10.1145/2896384.
[100] H. Sun, N. Ruan, and H. Liu, "Ethereum analysis via node clustering," in *Proc. NSS*, Taipei, Taiwan, 2019, pp. 147–158.
[101] H. Baek, J. Oh, C. Y. Kim, K. Lee, "A model for detecting cryptocurrency transactions with discernible purpose," in *Proc. ICUFN*, Zagreb, Croatia, pp. 713–717, 2019.
[102] R. S. Portnoff, D. Y. Huang, P. Doerfler, S. Afroz, D. McCoy, "Backpage and Bitcoin: Uncovering human traffickers," in *Proc. SIGKDD*, Halifax, NS, Canada, pp. 1595–1604, 2017.
[103] X. F. Liu, X.-J. Jiang, S.-H. Liu, C. K. Tse, "Knowledge discovery in cryptocurrency transactions: A survey," *IEEE Access*, vol. 9, pp. 37229–37254, 2021, doi: 10.1109/ACCESS.2021.3062652.
[104] A. Biryukov, S. Tikhomirov, "Deanonymization and linkability of cryptocurrency transactions based on network analysis," in *Proc. EuroS&P*, Stockholm, Sweden, pp. 172–184, 2019.
[105] C. Yu, C. Yang, Z. Che, L. Zhu, "Robust clustering of Ethereum transactions using time leakage from fixed nodes," *Blockchain: Res. Appl.*, vol. 4, issue 1, p. 100112, 2023, doi: 10.1016/j.bcra.2022.100112.
[106] A. Wahrstätter, J. Gomes, S. Khan, D. Svetinovic, "Improving cryptocurrency crime detection: CoinJoin community detection approach," *IEEE Trans. Dependable Secure Comput.*, 2023, doi: 10.1109/TDSC.2023.3238412.
[107] X. Zhao, Y. Jia, A. Li, R. Jiang, and Y. Song, "Multi-source knowledge fusion: A survey," *World Wide Web*, vol. 23, pp. 2567–2592, 2020, doi: 10.1007/s11280-020-00811-0.
[108] O. Elezaj, S. Y. Yayilgan, E. Kalemi, L. Wendelberg, M. Abomhara, and J. Ahmed, "Towards designing a knowledge graph-based framework for investigating and preventing crime on online social networks," in *Proc. e-Democracy*, Athens, Greece, 2019, pp. 147–162.
[109] S. Karur and P. S. Thilagam, "Crime Base: Towards building a knowledge base for crime entities and their relationships from online newspapers," *Inf. Process. Manag.*, vol. 56, no. 6, p. 102059, 2019, doi: 10.1016/j.ipm.2019.102059.
[110] M. Kejriwal and P. Szekely, "myDIG: Personalized illicit domain-specific knowledge discovery with no programming," *Future Internet*, vol. 11, no. 3, p. 59, 2019, doi: 10.3390/fi11030059.
[111] F. Carrillo-Brenes, L. M. Vilches-Blázquez, and F. Mata, "A proposal for semantic integration of crime data in Mexico City," in *Proc. GIS LATAM*, Mexico City, Mexico, 2020, pp. 25–38.
[112] H. Wu, "Research on electronic evidence management system based on knowledge graph," in *Proc. DependSys*, Guangzhou, China, 2019, pp. 369–380.
[113] Mazzonello et al., "The H2020 ANITA platform: Generating knowledge about crime through user-centred innovative tools," in *Proc. CSCI*, Las Vegas, NV, USA, 2021, pp. 679–684.
[114] W. Müller, D. Mühlenberg, D. Pallmer, U. Zeltmann, C. Ellmauer, and K. Demestichas, "Knowledge engineering and ontology for crime investigation," in *Proc. AIAI*, Crete, Greece, 2022, pp. 483–494.
[115] *State v. Loomis*, 881 N.W.2d 749 (Wis. 2016)
[116] W. D. Heave, "Predictive policing algorithms are racist. They need to be dismantled," *MIT Technol. Rev.*, July 17, 2020. [Online]. Available: https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/
[117] European Union. (2016, Apr. 27). *Regulation (EU) 2016/679 of the European Parliament and of the Council on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation)*. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj.
[118] European Parliament and Council. (2016, Apr. 27). *Directive (EU) 2016/680 on the Protection of Natural Persons with Regard to the Processing of Personal Data by Competent Authorities for the Purposes of the Prevention, Investigation, Detection, or Prosecution of Criminal Offences and on the Free Movement of Such Data*. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016L0680.
[119] European Union. (2024, Jun. 13). *Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence (AI Act)*. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2024/1689/oj.
[120] European Union. (2012, Oct. 26). *Charter of Fundamental Rights of the European Union*. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:12012P/TXT.
[121] European Commission. (2021). *Ethics By Design and Ethics of Use Approaches for Artificial Intelligence*. [Online]. Available: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf.
[122] J. Cheney-Lippold, *We are data: Algorithms and the making of our digital selves*, New York, NY, USA: NYU Press, 2017.
[123] P. Kaufmann, H. Kuch, C. Neuhaeuser, E. Webster, *Humiliation, degradation, dehumanisation: Human dignity violated*, Library of Ethics and Applied Philosophy, Springer Science & Business Media, 2011, p. 87.
[124] E. Finlay, "Autonomous weapon systems in armed conflict: Death by data, where is the dignity? An examination into the compatibility of AWS' algorithmic processes with human dignity," Master's thesis, University of Amsterdam, Amsterdam, Netherlands, 2022.
[125] European Commission, High-Level Expert Group on Artificial Intelligence. (2019). *Requirements of Trustworthy AI*. [Online]. Available: https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1.
[126] European Commission, High-Level Expert Group on Artificial Intelligence. (2020). *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment*. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.
[127] Information Commissioner's Office. (2023). *How to Use AI and Personal Data Appropriately and Lawfully*. [Online]. Available:

https://ico.org.uk/media/for-organisations/documents/4022261/how-to-use-ai-and-personal-data.pdf.
[128] D. Lyon (ed.), *Surveillance as social sorting: Privacy, risk and automated discrimination*, London, UK: Routledge, 2003.
[129] B. Schermer, "Risks of profiling and the limits of data protection law," in *Discrimination and Privacy in the Information Society*, B. Custers, T. Calders, B. Schermer, and T. Zarsky, Eds. Berlin, Heidelberg: Springer, 2013, vol. 3, pp. 145–167.
[130] V. Krotov, L. Johnson, L. Silva, "Tutorial: Legality and ethics of web scraping," *Commun. Assoc. Inf. Syst.*, vol. 47, doi: 10.17705/1CAIS.04724.
[131] A. Sachoulidou, C. Rego Oliveira, A. Kordoni, et al., "D8.2 – The project's ethical, data protection and social impact assessment," *TRACE Project*, 2024.
[132] H. Nissenbaum, *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford, CA, USA: Stanford Law Books, 2009.
[133] R. Finn, D. Wright, and M. Friedewald, "Seven types of privacy," in *European Data Protection: Coming of Age*, S. Gutwirth, R. Leenes, P. de Hert, and Y. Poullet, Eds. Dordrecht, Netherlands: Springer, 2013, pp. 3–32.
[134] C. Winter, R. V. Gundur, "Challenges in gaining ethical approval for sensitive digital social science studies," *Int. J. Soc. Res. Methodol.*, vol. 27, no. 1, pp. 31–46, 2022, doi: 10.1080/13645579.2022.2122226.
[135] G. Davies, "Shining a light on policing of the dark web: An analysis of UK investigatory powers," *J. Crim. Law*, vol. 84, no. 5, pp. 407–426, 2020, doi: 10.1177/0022018320952557.
[136] Autoriteit Persoonsgegevens, "Richtlijnen scraping door private organisaties en particulieren," 2024. Guidance available: https://www.autoriteitpersoonsgegevens.nl/uploads/2024-05/Handreiking%20scraping%20door%20particulieren%20en%20private%20organisaties.pdf
[137] Article 29 Data Protection Working Party, "Guidelines on data protection impact assessment (DPIA) and determining whether processing is 'likely to result in a high risk' for the purposes of Regulation 2016/679," European Commission, WP248 rev.01, Oct. 2017.
[138] European Commission, Directorate-General for Communications Networks, Content and Technology. (2019). *Ethics Guidelines for Trustworthy AI*. [Online]. Available: https://data.europa.eu/doi/10.2759/346720.
[139] ACM. (2018). *Code of Ethics and Professional Conduct*. [Online]. Available: https://www.acm.org/code-of-ethics.
[140] IEEE. (2019). *Ethically Aligned Design*. [Online]. Available: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf.
[141] C. Huang, Z. Zhang, B. Mao, X. Yao, "An overview of artificial intelligence ethics," *IEEE Trans. Artif. Intell.*, vol. 1, no. 1, pp. 1–21, 2022, doi: 10.1109/TAI.2022.3194503.
[142] D. Peters, K. Vold, D. Robinson, R. A. Calvo, "Responsible AI: Two frameworks for ethical design practice," *IEEE Trans. Technol. Soc.*, vol. 1, no. 1, pp. 34–47, March 2020, doi: 10.1109/TTS.2020.2974991.
[143] Y. Wu, "Ethically responsible and trustworthy autonomous systems for 6G," *IEEE Netw.*, vol. 36, no. 4, pp. 126–133, Aug. 2022, doi: 10.1109/MNET.005.2100711.
[144] M. Rizinski, H. Peshov, K. Mishev, L. T. Chitkushev, I. Vodenska, and D. Trajanov, "Ethically responsible machine learning in fintech," *IEEE Access*, vol. 10, pp. 97531–97554, 2022, doi: 10.1109/ACCESS.2022.3202889.
[145] D. Schiff, J. Borenstein, J. Biddle, and K. Laas, "AI ethics in the public, private, and NGO sectors: A review of a global document collection," *IEEE Trans. Technol. Soc.*, vol. 2, no. 1, pp. 31–42, Mar. 2021, doi: 10.1109/TTS.2021.3052127.
[146] N. Rodis, C. Sardianos, P. Radoglou-Grammatikis, P. Sarigiannidis, I. Varlamis, and G. T. Papadopoulos, "Multimodal explainable artificial intelligence: A comprehensive review of methodological advances and future research directions," *arXiv:2306.05731*, 2023.
[147] M. Hutson, "The opacity of artificial intelligence makes it hard to tell when decision-making is biased," *IEEE Spectr.*, vol. 58, no. 2, pp. 40–45, Feb. 2021, doi: 10.1109/MSPEC.2021.9340114.
[148] T. D. Huynh, N. Tsakalakis, A. Helal, S. Stalla-Bourdillon, and L. Moreau, "Explainability-by-design: A methodology to support explanations in decision-making systems," *arXiv:2206.06251*, 2022, doi: 10.48550/arXiv.2206.06251.
[149] A. Rai, "Explainable AI: From black box to glass box," *J. Acad. Mark. Sci.*, vol. 48, pp. 137–141, 2020, doi: 10.1007/s11747-019-00710-5.
[150] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. ICML*, Sydney, NSW, Australia, 2017, pp. 3145–3153.
[151] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. SIGKDD*, New York, NY, USA, pp. 1135–1144, 2016.
[152] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Proc. NIPS*, Red Hook, NY, USA, pp. 4768–4777, 2017.
[153] P. Schwab and W. Karlen, "CXPlain: Causal explanations for model interpretation under uncertainty," in *Proc. NIPS*, Red Hook, NY, USA, pp. 10220–10230, 2019.
[154] R. Hamon, H. Junklewitz, and J. Sanchez Martin, "Robustness and explainability of artificial intelligence," EUR 30040 EN, Publications Office of the European Union, Luxembourg, 2020.
[155] Y. Li et al., "A survey on dropout methods and experimental verification in recommendation," *IEEE Trans. Knowl. Data Eng.*, 2022, doi: 10.1109/TKDE.2022.3187013.
[156] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell, "Understanding data augmentation for classification: When to warp?" in *Proc. DICTA*, Gold Coast, QLD, Australia, 2016, pp. 1–6.
[157] V. Mygdalis and I. Pitas, "Hyperspherical class prototypes for adversarial robustness," *Pattern Recognit.*, vol. 125, p. 108527, 2022, doi: 10.1016/j.patcog.2022.108527.
[158] X. Ferrer, T. V. Nuenen, J. M. Such, M. Coté, and N. Criado, "Bias and discrimination in AI: A cross-disciplinary perspective," *IEEE Technol. Soc. Mag.*, vol. 40, no. 2, pp. 72–80, June 2021, doi: 10.1109/MTS.2021.3056293.
[159] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, and N. Shadbolt, "It's reducing a human being to a percentage: Perceptions of justice in algorithmic decisions," in *Proc. CHI*, New York, NY, USA, pp. 1–14, 2018.
[160] A. Visentin, A. Nardotto, and B. O'Sullivan, "Predicting judicial decisions: A statistically rigorous approach and a new ensemble classifier," in *Proc. ICTAI*, Portland, OR, USA, pp. 1820–1824, 2019.
[161] D. Pessach and E. Shmueli, "A review on fairness in machine learning," *ACM Comput. Surv.*, vol. 55, no. 3, 2023, doi: 10.1145/3494672.
[162] S. Verma and J. Rubin, "Fairness definitions explained," in *Proc. FairWare*, Gothenburg, Sweden, 2018, pp. 1–7.
[163] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," in *Proc. SIGKDD*, New York, NY, USA, pp. 797–806, 2017.
[164] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Weinberger, "On fairness and calibration," in *Proc. NIPS*, Long Beach, CA, USA, 2017, pp. 5680–5689.
[165] R. Stoykova, "Digital evidence: Unaddressed threats to fairness and the presumption of innocence," *Comput. Law Secur. Rev.*, vol. 42, p. 105575, 2021, doi: 10.1016/j.clsr.2021.105575.
[166] T. F. Blauth, O. J. Gstrein, and A. Zwitter, "Artificial intelligence crime: An overview of malicious use and abuse of AI," *IEEE Access*, vol. 10, pp. 77110–77122, 2022, doi: 10.1109/ACCESS.2022.3191790.
[167] D. Li, W. Wang, K. Zhao, J. Dong, and T. Tan, "RiDDLE: Reversible and diversified de-identification with latent encryptor," in *Proc. CVPR*, Vancouver, BC, Canada, 2023, pp. 8093–8102.
[168] Z. Bu, J. Mao, and S. Xu, "Scalable and efficient training of large convolutional neural networks with differential privacy," in *Proc. NIPS*, New Orleans, LA, USA, 2022, pp. 38305–38318.
[169] G. T. Papadopoulos, M. Antona, and C. Stephanidis, "Towards open and expandable cognitive AI architectures for large-scale multi-agent human-robot collaborative learning," *IEEE Access*, vol. 9, pp. 73890–73909, 2021, doi: 10.1109/ACCESS.2021.3080517.
[170] S. Thermos, G. T. Papadopoulos, P. Daras, and G. Potamianos, "Deep affordance-grounded sensorimotor object recognition," in *Proc. CVPR*, Honolulu, HI, USA, 2017, pp. 6167–6175.

Ioannis Mademlis (S'17-M'18-SM'22) is a computer scientist, specialized in artificial intelligence. He received a Ph.D. in machine learning and computer vision (2018) from the Aristotle University of Thessaloniki (AUTH), Greece. He was a postdoctoral research associate at AUTH (2018-'22) and at the Harokopio

University of Athens, Greece (2022-'24). In 2022-'23, he was an adjunct professor of machine learning at the Athens University of Economics and Business, Greece. He has participated in 6 European Union-funded R&D projects, having co-authored approximately 70 publications in academic journals and international conferences. His current research interests include machine learning, computer vision, autonomous robotics and human-computer interaction. He is a lecturer and a committee member of the EU-funded International Artificial Intelligence Doctoral Academy (IAIDA).

Marina Mancuso is a researcher at Transcrime, the Joint Research Centre on Innovation and Crime of the Università Cattolica del Sacro Cuore, the Alma Mater Studiorum Università di Bologna and the Università degli Studi di Perugia. She is also Adjunct Professor of Criminology at Università Cattolica del Sacro Cuore. She received an Honours Master's degree in Applied Social Sciences, specialising in Crime and Security and she graduated with the International PhD in Criminology, both at Università Cattolica del Sacro Cuore in Milan, Italy. She collaborated and coordinated different research projects at national and international level in the area of serious and organized crime and illicit markets.

Caterina Paternoster is a Ph.D. candidate in criminology at the Università Cattolica del Sacro Cuore of Milan. She is also researcher at Transcrime, the Joint Research Centre on Innovation and Crime of the Università Cattolica del Sacro Cuore, the Alma Mater Studiorum Università di Bologna and the Università degli Studi di Perugia. She holds a M.Sc. degree in Public Policies – curriculum Security Policies – at the Università Cattolica del Sacro Cuore of Milan in 2021. Her research interests include organized crime and the analysis of criminal networks and their structures.

Spyridon Evangelatos received the M.Eng. degree in electronics and radio-communications, the M.Sc. degree in signal processing for communications and multimedia, and the M.Sc. degree in control theory and computing from the National and Kapodistrian University of Athens (NKUA), in 2005, 2008, 2010, and 2013, respectively. He has served as a Research Fellow with the Self-Evolving Cognitive and Autonomic Networking Group, NKUA, and with the Physics of Information Laboratory, NKUA. He is currently a senior R&D expert for Netcompany-Intrasoft where he has been involved in several EU funded projects dealing with safety and security and also serves as a Research Associate with the Hellenic Mediterranean University (HMU). His current research interests include advanced signal processing techniques, Explainable AI and biometric technologies.

Emma Finlay received her BCL Law from University College Cork, Ireland (2021) and LLM International Law from Amsterdam Law School (2022). She is a research analyst in the Law Enforcement and Community Safeguarding cluster in Trilateral Research. She leads the legal, ethical, and societal work in the Horizon Europe project, CEASEFIRE. Her current research interests include the legal, ethical, and societal implications of AI-driven technology.

Joshua Hughes received LL.M (2014) and PhD (2020) degrees in international law from Lancaster University. He is currently a Research Manager and Cluster Lead for Law Enforcement and Community Safeguarding at Trilateral Research. He has led ethical analyses in 5 EU-funded research projects. His research interests include ethical, legal, and societal impacts of automation and AI in security technologies.

Panagiotis Radoglou-Grammatikis (Member, IEEE) received a M.Eng. and a Ph.D from the Department of Informatics and Telecommunications Engineering, University of Western Macedonia, Greece, in 2016 and 2023, respectively. His main research interests are in the area of cybersecurity and mainly focus on cyber-AI, intrusion detection and security games. He has published more than 30 research papers in international scientific journals, conferences and book chapters. He participates in the Topical Advisory Panel of Electronics (MDPI Publishing) and he is working as an R&D director at K3Y Ltd, coordinating the technical activities and strategy of K3Y in various R&D projects. Moreover, he is co-founder of MetaMind Innovations P.C., the first spin-off of the University of Western Macedonia. He is a member of ACM and the Technical Chamber of Greece.

Panagiotis Sarigiannidis (Member, IEEE) received the B.Sc. and Ph.D. degrees in computer science from the Aristotle University of Thessaloniki, Greece, in 2001 and 2007, respectively. His research interests include telecommunication networks, the Internet of Things, and network security. He is the Director of the ITHACA Laboratory (https://ithaca.ece.uowm.gr/), Kozani, Greece, a co-founder of the first spin-off of the University of Western Macedonia (MetaMind Innovations P.C.) and an Associate Professor with the Department of Electrical and Computer Engineering, University of Western Macedonia, Kozani, Greece. He has published over 270 papers in international journals, conferences, and book chapters. He participates on the editorial boards of various academic journals.

Georgios Stavropoulos received the Diploma degree in Electrical and Computer Engineering from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2006. He is a research associate with the Information Technologies Institute of the Centre for Research and Technology Hellas (CERTH/ITI) as well as Chief Technology Officer of Parhelia Analytics Ltd. (CERTH/ITI Spin-off Company). His main research interests include signal processing, computer vision, machine learning, visual analytics, and more. Since 2006, he has been involved in numerous European and National projects, as an expert developer and/or technical manager. He has co-authored more than 30 articles in refereed journals and international conferences. He is a member of the Technical Chamber of Greece.

Konstantinos Votis received an MSc and a Ph.D. degree in computer science and service-oriented architectures from the University of Patras, Greece. HE also holds an MBA from the Business School department in the University of Patras. Presently, he is a computer engineer and a senior researcher (Researcher Grade B') at Information Technologies Institute/Centre for Research and Technologies Hellas (CERTH/ITI) and Director of the Visual Analytics Laboratory of CERTH/ITI. He is also a visiting professor in the University of Nicosia, Institute of the Future, regarding Blockchain and AI technologies (since October 2019). He was also a Visiting professor at the De Montfort University in UK in the field of Human Computer Interaction, Virtual and Augmented Reality (2016-2020). His research interests include Human Computer Interaction (HCI), information visualization and management of big data, knowledge engineering and decision support systems.

Georgios Th. Papadopoulos (S'08–M'11) received the M.Eng. and Ph.D. degrees in electrical and computer engineering from the Aristotle University of Thessaloniki, Greece, in 2005 and 2011, respectively. He is currently an Assistant Professor at the Department of Informatics and Telematics of the Harokopio University of Athens, Greece. His research interests include computer vision, machine/deep learning and artificial intelligence. He has more than 70 publications in international academic conferences and scientific journals. He has over 20 years of experience in participating to EU-funded R&I projects in the areas of ICT, security and robotics. He is a member of the Technical Chamber of Greece.