electronics

MDPI

Opinion

Towards an AI-Enhanced Cyber Threat Intelligence Processing Pipeline

Lampis Alevizos [1,*] and Martijn Dekker [2]

[1] School of Engineering and Computer Science, University of Central Lancashire (UCLan), Preston PR1 2HE, UK
[2] Faculty of Economics and Business, Amsterdam Business School, University of Amsterdam (UvA), Amsterdam 1018 TV, The Netherlands; m.dekker4@uva.nl
* Correspondence: lampis@redisni.org

Abstract: Cyber threats continue to evolve in complexity, thereby traditional cyber threat intelligence (CTI) methods struggle to keep pace. AI offers a potential solution, automating and enhancing various tasks, from data ingestion to resilience verification. This paper explores the potential of integrating artificial intelligence (AI) into CTI. We provide a blueprint of an AI-enhanced CTI processing pipeline and detail its components and functionalities. The pipeline highlights the collaboration between AI and human expertise, which is necessary to produce timely and high-fidelity cyber threat intelligence. We also explore the automated generation of mitigation recommendations, harnessing AI's capabilities to provide real-time, contextual, and predictive insights. However, the integration of AI into CTI is not without its challenges. Thereby, we discuss the ethical dilemmas, potential biases, and the imperative for transparency in AI-driven decisions. We address the need for data privacy, consent mechanisms, and the potential misuse of technology. Moreover, we highlight the importance of addressing biases both during CTI analysis and within AI models, warranting their transparency and interpretability. Lastly, our work points out future research directions, such as the exploration of advanced AI models to augment cyber defenses, and human–AI collaboration optimization. Ultimately, the fusion of AI with CTI appears to hold significant potential in the cybersecurity domain.

Keywords: artificial intelligence; cyber threat intelligence; cyber resilience; ethical considerations; CTI and AI biases

Citation: Alevizos, L.; Dekker, M. Towards an AI-Enhanced Cyber Threat Intelligence Processing Pipeline. *Electronics* **2024**, *13*, 2021. https://doi.org/10.3390/electronics13112021

Academic Editors: Juan-Carlos Cano and Aryya Gangopadhyay

Received: 31 March 2024
Revised: 12 May 2024
Accepted: 16 May 2024
Published: 22 May 2024

1. Introduction and Motivation

Cyber threats are continuously growing in complexity and frequency, therefore the ability to rapidly process and act upon cyber threat intelligence (CTI) can mean the difference between a mitigated threat and a breach. CTI, as defined by the National Institute of Standards and Technology (NIST), includes information that allows organizations to understand the latest threats and to proactively defend against them [1]. However, the vast volume of CTI, coupled with its dynamic nature, poses significant challenges for timely processing and action.

Traditional CTI processing methodologies involve manual efforts, where analysts examine large amounts of data, attempting to recognize patterns, validate intelligence, and recommend actions [2]. Namely, analysts are trying to produce actionable and valuable intelligence by contextualizing information. This manual approach, while valuable, is increasingly becoming unsustainable given the scale and speed of modern cyber threats. The need for automation and enhanced analytical capabilities, therefore, has become evident.

AI, with its ability to handle large datasets and its capability to learn and adapt, offers a promising direction to augment the CTI processing pipeline. Preliminary research, such as the work by Buczak and Guven [3] have has already highlighted the potential of AI in

cybersecurity, particularly in areas like anomaly detection and malware classification. However, the integration of AI into the CTI processing pipeline, especially in a manner that highlights the collaboration with human expertise, remains an area of research.

This paper seeks to bridge this gap, presenting a comprehensive approach to harnessing AI for CTI processing. Our focus goes beyond automation, creating a collaborative framework where AI and human analysts work together, to produce rapid, accurate, and actionable CTI. By streamlining this pipeline, we aim to reduce the time from intelligence ingestion to the implementation of mitigating measures and, subsequent, resilience verification. We believe that combining AI with CTI offers a proactive and adaptable cybersecurity approach, rather than a reactive one. Our goal is to connect AI's capabilities with cybersecurity requirements, advancing future innovations in the field. The contributions by this paper can be summarized as follows:

(1) A blueprint of the AI-enhanced CTI processing pipeline: We present a comprehensive framework that integrates AI techniques at various stages of a threat-informed defense, starting with CTI data ingestion and progressing to resilience verification. We detail the components and functionalities, as well as highlighting the imperative collaboration between AI and human expertise.
(2) Innovation in real-time and predictive threat mitigation: Our research pioneers the use of AI for generating real-time, contextual, and predictive mitigation strategies. We explore the application of advanced AI algorithms that can swiftly analyze CTI data and suggest security measures, thereby enhancing the organizational responsiveness to cyber threats.
(3) Ethical and bias considerations: We perform a thorough examination of the ethical implications of using AI in CTI and strategies to address potential biases in both the CTI domain and AI models. We also propose methods to ensure unbiased and transparent AI-driven insights.
(4) Introduction of a cyber resilience index: We propose a novel cyber resilience index that serves as a barometer for an organization's defensive capabilities against cyber threats. Analogous to financial market indices, this metric offers a quick overview of an organization's cyber health, informing strategic defense decisions.
(5) Challenges in AI-driven CTI: We critically discuss the hurdles to embedding AI into CTI, starting with the ethical dilemmas, data bias, and the need for transparency in AI-driven decisions, presenting a roadmap for addressing these issues.
(6) Future research directions: We provide future research directions emerging from our findings, thus underlining areas of potential growth and innovation.

The structure of this paper is as follows. We begin with the background and a literature review. Next, we detail the components of the AI-enhanced CTI processing pipeline, namely (A) intelligence ingestion, (B) collaborative analysis, (C) automated mitigation, and (D) resilience verification. In the next section, we discuss the challenges and considerations for the AI-enhanced CTI processing pipeline. Lastly, we summarize the conclusions and propose future research directions.

2. Background and Literature Review

The convergence of CTI and AI is a relatively emerging field, but one that has gained significant attention due to the potential benefits it promises. CTI has evolved from basic threat feeds to sophisticated intelligence platforms that provide contextual information about threats [4]. The primary goal of a mature CTI capability should be to offer actionable and valuable insights that can guide defensive measures, essentially extracting the right signals throughout the vast "noise" within the cyber landscape [5]. The works of Chen et al. [6] provide a comprehensive overview of the CTI landscape, highlighting the challenges associated with intelligence validation and relevance determination.

The application of AI in cybersecurity is not completely new. Machine learning models have been employed for tasks like spam detection and network intrusion detection for years, as detailed in the work of Sarker et al. [7]. However, the integration of AI with CTI

is a more recent endeavor and lacks research output. The potential of AI to process vast amounts of data rapidly makes it a natural fit for CTI processing. For instance, Ring et al. [8] explored the use of AI for threat hunting, highlighting the potential to uncover hidden threats in vast datasets.

Despite the advancements, several challenges persist in regard to CTI processing. The dynamic nature of the cyber threat landscape, coupled with the large volume of data, often leads to information overload [9]. Additionally, false positives, outdated intelligence, and a lack of context can hamper the effectiveness of CTI. Sauerwein et al. [10] researched these challenges, offering insights into potential mitigation strategies.

The collaboration between AI and human expertise is also an important topic of considerable interest. While AI excels at processing large datasets, human intuition and expertise remain irreplaceable for nuanced threat analysis [11]. The challenge lies in creating a framework or pipeline where AI augments human capabilities, without overwhelming them with data. Brundage et al. [12] discussed the potential pitfalls and best practices for AI–human collaboration in decision-making contexts. In our work, we aim to bridge this gap by establishing strong collaborative bonds between the CTI analyst and AI, where both complement each other's strengths and counter each other's weaknesses.

The work of Varma et al. [13] work primarily targets small and medium-sized enterprises (SMEs), providing them with a roadmap to integrate AI into their CTI processes. Although the work offers valuable insights for SMEs, its scope is limited when considering larger organizations, or more complex cybersecurity infrastructure. The researchers used AI only to enhance CTI, rather than create a comprehensive solution for cybersecurity. Our paper on the other hand, presents a comprehensive AI-enhanced CTI processing pipeline that is adaptable to organizations of any size. Additionally, the roadmap in the referenced work can be seen as a practical application, while our proposed pipeline offers a broader perspective, including all stages of CTI processing, making it more universally applicable, while also allowing organizations to select the components that best suit their needs. Suryotrisongko et al. [14] underlined the importance of trust in CTI sharing, advocating for the blending of explainable AI (XAI) with open-source intelligence (OSINT). While the work underlines the significance of transparency in AI-driven CTI, it primarily focuses on botnet detection, which is a specific subset of the broader CTI landscape. Our focus on transparency and interpretability in AI-driven insights aligns with the principles of XAI. However, our goal is to provide a more holistic view of the CTI landscape, addressing various challenges and stages of CTI processing, beyond just botnet detection. Ranade et al. [15] studied a niche, but crucial, challenge in regard to CTI, namely the generation of fake CTI descriptions using advanced AI models. Although the work highlights an emerging threat in the CTI domain, its primary focus was on the generation aspect rather than mitigation or validation. In our work we highlight the importance of validation and relevance determination in CTI. The challenge of fake CTI generation further highlights the need for robust validation mechanisms, which our paper addresses in detail, offering solutions and strategies to counter such threats.

Moraliyag et al. [16] proposed a proactive approach to CTI by classifying onion services based on content. Onion services, also known as hidden services, are services that are hosted on the Tor network (https://www.torproject.org/ (accessed on 4 January 2024)). Unlike traditional websites with public IP addresses, onion services use the Tor network's anonymizing technology to protect both users and service operators. Although this work provides valuable insights into dark web intelligence, its primary focus remains on classification techniques, potentially overlooking other crucial aspects of CTI processing. Our intelligence ingestion phase, in the proposed pipeline, can benefit from such classification techniques. Nonetheless, we propose a more comprehensive view, detailing various stages of CTI processing and addressing challenges beyond just classification. The research by Mitra et al. [17] research focuses on enhancing cybersecurity knowledge graphs with intelligence provenance, which is a novel approach to combat fake CTI. Nevertheless, relying solely on provenance might not address all the challenges associated with fake

CTI, especially when considering sophisticated adversarial attacks. The integration of provenance information aligns with our pipeline's intelligence ingestion and collaborative analysis phases. However, in this work we provide a multi-faceted approach to CTI validation, which enables a more robust form of defense against fake intelligence. Mittal et al. [18] discussed the potential of AI in CTI, highlighting its role in uncovering hidden threats. Whilst this work offers valuable insights, it does not provide a detailed roadmap or framework for integrating AI into CTI processing. Our work builds on this premise but goes a step further by detailing how AI can be systematically integrated into various stages of CTI processing, offering a more structured approach.

3. The AI-Enhanced CTI Processing Pipeline

The fusion of AI capabilities with CTI processing has the potential to significantly enhance the speed, accuracy, and efficiency of threat intelligence operations. This section outlines a structured pipeline that integrates AI at various stages of CTI processing to enable the abovementioned attributes. Figure 1 visualizes the individual components comprising the AI-enhanced CTI processing pipeline, as a blueprint.

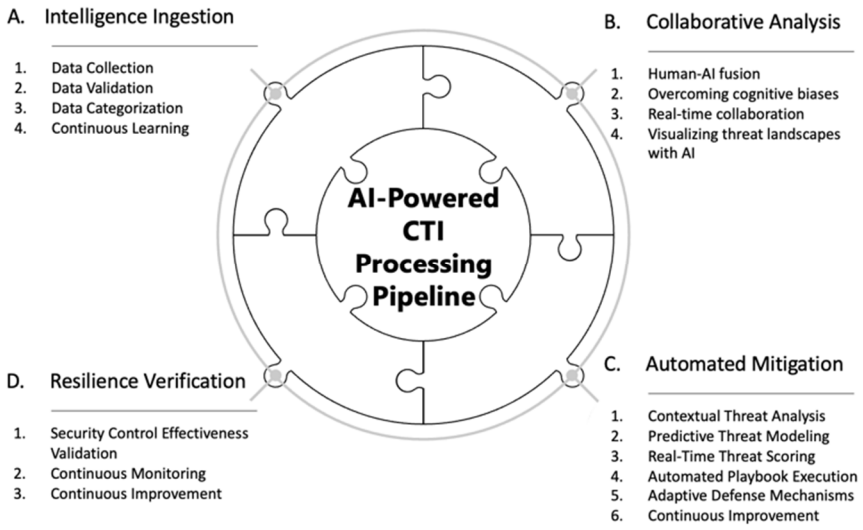

Figure 1. AI-enhanced CTI processing pipeline blueprint.

The central theme of Figure 1 is represented by the middle circle labelled "AI-enhanced CTI Processing Pipeline". Radiating outward from this central theme are four main components in a puzzle shape, namely:

A. Intelligence ingestion, which focuses on the initial stages of data collection, data validation, and data categorization using AI;
B. Collaborative analysis, which focuses on the collaborative analysis between human intelligence and artificial intelligence. We detail the concept of human–AI fusion, ways of overcoming cognitive biases, real-time collaboration, and visualizing threat landscapes with AI;
C. Automated mitigation, which focuses on analyzing threats in context using AI. It contains predictive threat modelling, real-time threat scoring, automated playbook execution, adaptive defense mechanisms, and a feedback loop for continuous improvement;
D. Resilience verification, which comprises of a proactive approach to security by simulating cyber-attacks. The focus is on continuous monitoring and continuous improvement led by AI, ultimately leading to a single cyber resilience metric, the cyber resilience index.

Each of these main components breaks down further into subcomponents, which we detail in the corresponding sections.

3.1. Intelligence Ingestion

The initial phase of the CTI processing pipeline is the ingestion of raw information or intelligence data. This means collecting, validating, and categorizing vast amounts of data from various sources, such as threat feeds, logs, and other intelligence or information repositories. The threat landscape is dynamic, with new threats emerging regularly; thereby, it is imperative for a CTI ingestion process, empowered by AI, to continuously learn from new data and adapt to the evolving threat environment. Figure 2 outlines the intelligence ingestion steps.

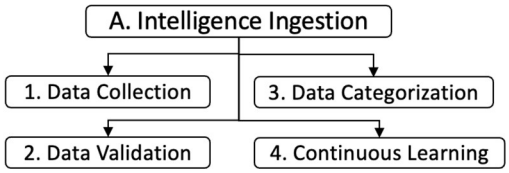

Figure 2. Intelligence ingestion steps.

3.1.1. Data Collection

Given the vast amount of CTI data that needs to be collected, manual collection, although feasible, is extremely time consuming and the added value of such an exercise is highly dependent on the analysts experience and expertise alone. AI-driven tools orchestrate multiple APIs (Application Programming Interfacesapplication programming interfaces) and, therefore, automate the collection process; thus, data is gathered in real-time and from a wide range of sources, such as open-source intelligence (OSINT), human intelligence (HUMINT), dark web monitoring, commercial threat feeds, internal organization threat data, vendor reports, social media monitoring, threat intelligence platforms (TIPs), and industry-specific threat reports.

3.1.2. Data Validation

Not all collected data are relevant. AI algorithms can quickly examine the data, discarding irrelevant information and highlighting potential threats using decision trees, as shown by Kotsiantis [19]. Then, such algorithms cross-reference data from multiple sources to verify its accuracy, using graph analytics to map the relationships between sources. For instance, if two independent sources report the same threat, it is more likely to be credible. However, the challenge is not just about collecting data, but producing meaningful and actionable intelligence from the overwhelming amount of "noise". Transforming information into actionable, valuable intelligence should be the goal. The total volume of data, combined with its dynamic nature, makes manual validation and analysis a challenging task. This is where AI plays a transformative role and can be implemented based on the following elements.

Signal extraction, using convolutional neural networks (CNNs), provides proven pattern recognition within large datasets, enabling such networks to detect indirect signs of malicious activity that may signify cyber threats. This capability allows AI to effectively extract the "signal", the meaningful, actionable intelligence, from the "noise", irrelevant or redundant information. Goodfellow et al. [20] provided the foundational theory on the capabilities of neural networks in terms of pattern recognition and anomaly detection. TensorFlow v2.16.0 (https://www.tensorflow.org/tutorials/generative/autoencoder, accessed on 12 December 2023) serves as a practical implementation of that theory, being a popular open-source machine learning framework that offers long short-term memory (LSTM) autoencoders, which can be used for anomaly detection in time-series data [21].

Cross-referencing and identifying correlations utilizing AI tools from multiple data sources automatically, enhances the validation process. For instance, if two independent sources report the same threat, the AI assigns a higher credibility score to that piece of

intelligence. Advanced algorithms can also correlate seemingly unrelated pieces of information, uncovering hidden threats or tactics used by adversaries. A prime theoretical example, provided by Landwehr et al. [22] introduced logistic model trees, which can be used for correlating and cross-referencing data from multiple sources. Elastic Stack (https://www.elastic.co/, accessed on 19 December 2023) (Elasticsearch, Logstash, Kibana) is a practical implementation of this, which is widely used in cybersecurity for its capabilities in terms of data ingestion, indexing, and visualization. It can correlate logs from various sources to provide a unified view of events.

Transforming information into intelligence: AI bridges this gap by analyzing the context in which data are generated, determining its relevance to the organization's threat landscape, and providing actionable recommendations based on the analyzed intelligence [23]. However, it is important to note that the effectiveness of AI depends upon the accuracy and quality of the underlying data, e.g., data within the configuration management database (CMDB) of a company. The quality and integrity of the data sources are therefore crucial, as they directly impact the reliability of the intelligence generated.

Addressing human limitations: Traditional CTI relies on human analysts who, despite their expertise, have limitations in terms of processing capacity and speed. AI complements human analysts by managing vast datasets, ensuring that no potential threat goes unnoticed [3]. This collaboration between human intuition and AI's computational aptitude provides for comprehensive threat intelligence.

A feedback mechanism exists, where false positives or irrelevant data flagged by AI are reviewed and fed back into the system for continuous learning and improvement. This iterative process improves the AI model's accuracy over time. Figure 3 demonstrates the process of AI-driven validation, and the transformation of raw data into actionable intelligence.

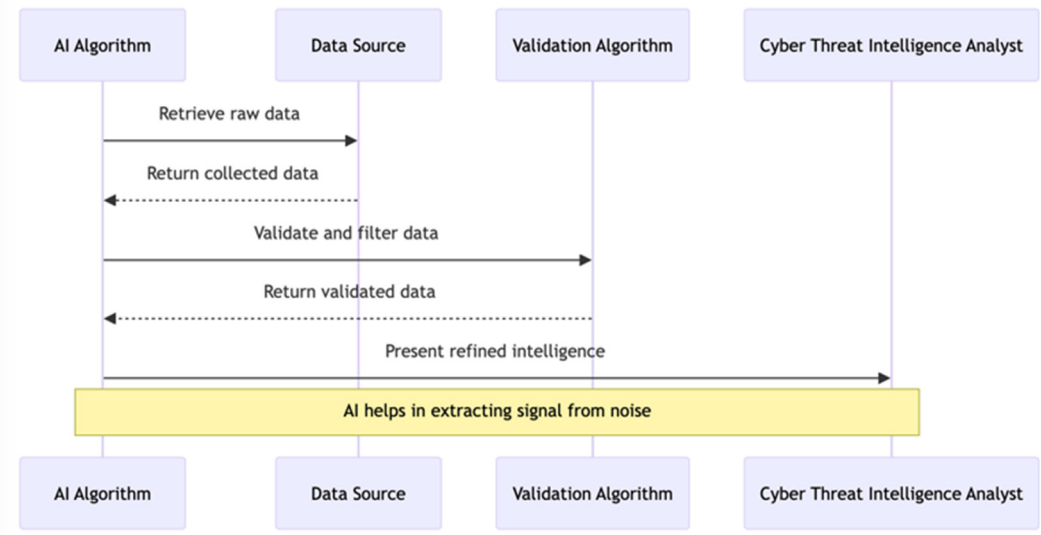

Figure 3. Extracting the signal from the noise using AI.

The AI algorithm retrieves raw data from the data source. The collected data is returned to the AI algorithm. Next, the AI algorithm sends this data to the validation algorithm to validate and filter out irrelevant or noisy data. The validated data is then returned to the AI algorithm. Finally, the refined intelligence is presented to the CTI analyst.

3.1.3. Data Categorization

Effective CTI requires the ingested data to be categorized into meaningful segments, which can guide the subsequent analysis and action as a result of the human–AI collaboration. The Latent Dirichlet Allocation (LDA) algorithm can be used to identify the underlying themes or topics within large text datasets, helping to categorize the content by subject matter [24]. For instance, threat actors (TAs), threat events, TTPs (tactics, techniques, and procedures), indicators of compromise (IoCs), the goals and motivations of TAs, and geopolitical trends. Moreover, Schonlau et al. showed how to use the random forest algorithm in this regard [25], while Sarker showed how to use neural networks to classify data into predefined categories based on training datasets [26], where the classification criteria are already known. Thus, using either method, AI takes unstructured data and organizes it into meaningful segments, ready for further analysis and action in the CTI pipeline.

3.1.4. Continuous Learning

The cyber threat landscape is not static; it evolves continuously with new vulnerabilities, TTPs, and threat actors emerging regularly. For an AI-enhanced CTI pipeline to remain effective, it must adapt to these changes. To successfully enable the AI-enhanced CTI pipeline to continuous learning, several methods can be used.

Machine learning (ML) for continuous relevance in CTI can be achieved utilizing the online learning algorithm [27]. This algorithm incrementally updates parameters in response to each new data point, thus providing adaptability to emerging threats without full retraining. This approach keeps the predictive model current, according to the evolving threat landscape.

Adversarial machine learning (AML) is used to anticipate potential evasion techniques that adversaries might employ. Red teams can either perform traditional attack simulations or use AML to simulate advanced evasion tactics. This will result in collecting data on novel attack vectors and improving defenses before they are exploited in the wild. Red teams should create adversarial examples led by cyber threat intelligence to assess an organization's defenses. Any successful evasion that is logged and analyzed, is in turn used to improve the data collection mechanisms. As the AI system processes new threat intelligence and interacts with human analysts, it will inevitably encounter false positives or misclassifications. Therefore, incorporating feedback from these interactions will allow the system to refine its algorithms, reducing errors over time.

Defensive distillation should be used to make machine learning models more robust against adversarial attacks [28]. Therefore, the data being collected will not be polluted by adversarial noise. Leveraging this technique to train the model on a "softened" version of the data, where the output probabilities are smoothed or "distilled" will make it harder for adversaries to find the precise distresses needed to deceive the model, thus data collection is cleaner.

Incorporating the broader context is important, which means that the AI system should continuously learn and incorporate insights from broader geopolitical, technological, and socio-economic contexts, enhancing its threat predictions. Political tensions often correlate with targeted cyber-attacks on government and critical infrastructure. Monitoring such geopolitical developments will allow the pipeline to anticipate increased risk levels and identify potential aggressors. Moreover, technological trends impact the nature and prevalence of cyber threats, if an innovative technology becomes widespread (e.g., blockchain), the AI system will prioritize threats targeting that technology. Lastly, economic challenges often lead to increased cybercrime activity. Tracking socio-economic indicators will help anticipate a rise in certain threat types. For instance, during economic downturns, there is typically a rise in financial fraud schemes.

Active learning is a specialized form of machine learning, where the model actively queries the human analyst for inputs on specific predictions [29]. For instance, if the AI system

encounters a piece of data it is uncertain about, it seeks confirmation from a human expert. Over time, these interactions reduce the system's uncertainty and improve its accuracy.

3.2. Collaborative Analysis

Traditional analysis methods relying heavily on human expertise are oftentimes slow, potentially biased, and prone to errors. This is where artificial intelligence and, more specifically, machine learning, can play a crucial role [30]. Mishra's work [31] showed that gradient boosting machines (GBMs), trained on historical threat data, can provide insights into potential threats, their patterns, and possible implications. Human analysts collaborate with these AI insights, leveraging their expertise to understand the nuances and context behind each threat. Figure 4 outlines the collaborative analysis steps.

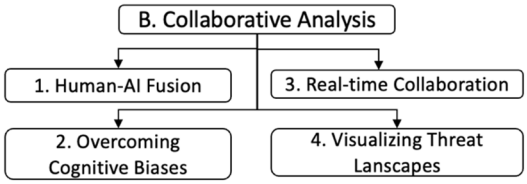

Figure 4. Collaborative analysis steps.

3.2.1. The Human–AI fusion

The collaboration between human analysts and AI is not about replacing one with the other, but about amplifying the strengths and mitigating the weaknesses of both. A summary of this phase is visualized in Figure 5.

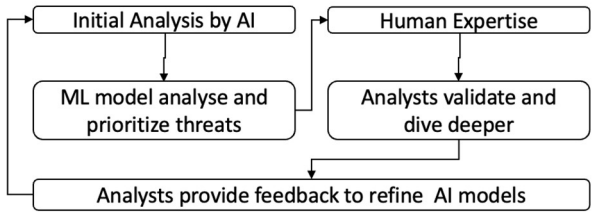

Figure 5. Human–AI Fusion.

The initial analysis by AI provides speed and scale. AI processes vast amounts of data at speeds incomprehensible to humans. For instance, Desmond et al. [32] showed that a model trained on extensive datasets can analyze millions of logs within minutes to detect anomalies, as opposed to humans. Chen's work [33] showed that AI can also provide pattern recognition through a deep learning algorithm, which is able to recognize patterns in data. Goodfellow et al. [20] proposed the use of neural networks that can identify patterns associated with malware traffic in network logs, even if the malware is a zero-day variant. In addition, AI offers fast prioritization. Based on detected patterns and historical data, AI can prioritize threats; therefore, the most imminent and dangerous threats can be addressed first. Existing AI-driven tools, like Cisco's Threat Grid v2.15 (https://www.cisco.com/c/en/us/products/security/threat-grid/index.html, accessed on 4 January 2024), analyze millions of samples daily, providing automated threat scores, based on the potential impact and prevalence of the detected threats [34].

Human expertise provides a broader contextual understanding. Although AI can recognize patterns, human analysts understand the broader context. For instance, in the SolarWinds attack (https://www.wired.com/story/the-untold-story-of-solarwinds-the-boldest-supply-chain-hack-ever/, accessed on 4 January 2024), while AI tools detected anomalies, it was the human analysts who pieced together the broader campaign, understanding the implications and the actors behind it. Human analysts bring intuition and experience. Analysts, with years of professional experience, can subjectively understand when

something does not seem right, even if it passes AI checks. Their expertise allows them to focus on complex threats specific to the IT landscape, forming hypothesis and, therefore, uncovering potential hidden connections. Lastly, human analysts play a crucial role in validating AI findings. Although AI might flag a potential phishing email based on certain patterns, a human analyst can validate it by considering the sender's context, the email's content, and other information. Analysts interacting with the AI-enhanced pipeline must provide feedback for refining the AI model, thus its predictions become more accurate over time. However, a challenge to using AI in cybersecurity is the potential for false positives [10]. Human analysts should flag something as a false positive, so that eventually, the AI learns from it, thereby reducing similar false alarms in the future. False positives in AI cybersecurity systems are problematic because they waste the analyst's time and resources, leading to alert fatigue and potential oversight of genuine threats. By learning from flagged false positives, AI models can adapt and improve their detection accuracy over time. Ultimately, this will reduce unnecessary alerts and allow analysts to remain focused on real threats, while the AI-enhanced CTI pipeline increases its trustworthiness over time.

3.2.2. Overcoming Cognitive Bias

Cognitive biases are systematic patterns of deviation from the norm or rational judgment, thus leading analysts to create their own subjective reality from their perception of the input [35]. Such biases can significantly impact the decisions of CTI analysts, thereby potentially leading to disregarded threats or data misinterpretations.

One of the major ransomware-related attacks (https://www.csoonline.com/article/563017/wannacry-explained-a-perfect-ransomware-storm.html, accessed on 4 January 2024) happened in 2017. In the aftermath of this case, many organizations focused heavily on protecting themselves against similar ransomware threats. Although this is a valid concern, an overemphasis on one type of threat due to its recent occurrence (availability heuristic) can lead to neglecting other potential threats. The AI-driven CTI pipeline provides for a balanced focus on all the relevant threats, not just those that are currently in the spotlight. The integration of AI into the CTI process offers a unique opportunity to counteract these biases, ultimately allowing for more objective and comprehensive analysis.

CTI analysts may be subject to the following biases:

A. Confirmation bias: the analyst might prioritize data that aligns with their existing threat models, potentially overlooking new or unexpected threats. The analyst's perspective is unintentionally influencing the collection, analysis, and interpretation of CTI data. This bias (also known as observer bias) can lead analysts to favor information that confirms their presumptions or to overlook data that contradicts their beliefs, impacting the accuracy and objectivity of their analysis;
B. Availability heuristic: the analyst might give undue weight to a recent high-profile cyber-attack, neglecting other potential threats;
C. Anchoring bias: the analyst relies too heavily on the first piece of information encountered (the "anchor") when making decisions. Oftentimes, CTI analysts anchor their analysis directly to initial findings, therefore missing the broader threat landscape;
D. Status quo bias: a preference for the current situation, resisting change, leading to an over reliance on established threat models and an inability to adapt to the evolving cyber threat landscape.

AI's role in mitigating biases:

A. AI algorithms are inherently decoupled from emotions and preconceived notions, thereby providing an objective analysis of data. They treat each piece of information based on its merits and relevance, not on any external influence or bias [36];
B. AI models apply consistent criteria when analyzing data, thus using the same standards across all data, contrary to the human analyst, who might unconsciously alter their criteria based on their biases [37];

C. AI analyses data for decision making based on comprehensive data, rather than anecdotal evidence or recent events [37].

In conclusion, to successfully overcome cognitive bias, the goal is not to replace human analysts with AI, but to have them work together. AI provides objective analysis, empowered by human analysts, who bring contextual understanding and intuition. By working together, they can counteract the biases inherent in both human judgment and AI models, therefore leading to more balanced and comprehensive threat analysis.

In exploring how humans and AI work together to counteract bias and optimize collaboration, Dell'Acqua et al. [38] highlighted two main ways: the 'Cyborg' and the 'Centaur.' The 'Cyborg' mode mixes human and AI efforts closely, using AI for its fast processing and humans for their deep understanding and moral judgment. On the other hand, the 'Centaur' mode is about humans and AI working side by side, with each taking on tasks that suit their strengths. This division helps make the most of both AI's data handling abilities and human creativity and ethical insights. These dimensions can serve as guardrails and show how combining human and AI strengths can enhance strategic decision making (centaur) and improve efficiency and accuracy (cyborg).

3.2.3. Real-Time Collaboration

Real-time machine learning models bring a change in thinking in regard to how threat intelligence is processed and acted upon. Consider a zero-day vulnerability that has just been discovered, traditional threat intelligence systems might take hours, if not days, to update their databases and provide recommendations. However, a real-time AI-driven system can pick up discussions about this vulnerability from sources like social media, forums, commercial tools, TIPs, or dark web marketplaces within minutes. Such a system can then assess the potential impact of this vulnerability, generate alerts for human analysts, and even recommend immediate countermeasures [39]. The successful real-time collaboration between the CTI analyst and AI is based on the following elements:

A. Dynamic data ingestion, as cyber threat data is generated by all of the above-described sources continuously, it is imperative to have a system that can ingest this data in real-time. AI-driven models, especially those built on streaming data platforms, can process data as it flows in, without waiting for batch updates [40];
B. Instantaneous analysis, once the data are ingested. Real-time machine learning models analyze data instantaneously [41]. This means that as soon as a new threat indicator is detected, the AI-enhanced CTI pipeline can assess its severity, potential impact, and relevance;
C. Real-time alerts based on instantaneous analysis. The AI-enabled CTI pipeline generates real-time alerts for human analysts. These alerts can be prioritized based on the potential impact and, thereby, the analysts can focus on the most pressing threats first;
D. Human–AI interaction; real-time collaboration should not be just one way. Human analysts, upon receiving alerts, can interact with the AI-enhanced CTI pipeline, asking follow-up questions or clarifications;
E. Adaptive learning; one of the differentiating factors of real-time machine learning models is their ability to learn on-the-fly. As new data is processed, the model can update its understanding, hence its predictions and recommendations are always based on the latest threat intelligence [42].

3.2.4. Visualizing Threat Landscapes with AI

The ability to visualize and quickly comprehend threat models is paramount. As an example, one could think of a scenario where a CTI analyst comes across an image detailing the flow of a sophisticated malware attack. Instead of spending hours, or even days, deciphering the image, the analyst can use an AI tool to quickly understand the malware's propagation, its potential targets, and its behavior. The AI tool can also cross-reference the

malware's signature with a database, providing insights into its origin, past variants, and potential countermeasures [43]. There are two prime examples where AI can supercharge collaborative analysis and empower the CTI analyst, as follows:

(i). Automated threat modelling:

Traditional threat modelling is a time-consuming exercise, requiring analysts to manually map out threats, vulnerabilities, and potential attack vectors. Moreover, CTI analysts and relevant stakeholders may lack the technical knowledge to perform threat modelling. Akhtar et al. [44] showed how AI automates this process, rapidly creating threat models based on the available data. Moreover, as new threat intelligence is ingested, AI can dynamically update the threat model, thus always reflecting the current threat landscape. Lastly, AI algorithms and especially deep learning models can be used to identify difficult correlations and potential threats that might be overlooked in manual analysis [45].

(ii). Image recognition and explanation:

CTI analysts can drop images depicting complex IT landscapes, or complex threat actor flows, into an AI-powered tool. Iqbal et al. showed how AI can instantly analyze the image, recognizing various components, connections, and potential vulnerabilities [7]. Furthermore, in the same work, it was proven that AI can go beyond simple recognition. Advanced AI models can provide contextual explanations for the elements in the image [7]. For instance, if an analyst drops an image of a network topology, the AI can identify servers, firewalls, potential choke points, and even suggest potential attack vectors based on the layout. Furthermore, by cross-referencing the elements in the image with historical threat data, AI can provide insights into past vulnerabilities, attacks, or breaches associated with similar setups [45].

3.3. Automated Mitigation

Based on the combined intelligence from AI and human analysis, the pipeline integrates with organizational tools like configuration management databases (CMDBs), or taps into IT infrastructure data, to understand the environment and adapt recommendations. The recommendations range from technical solutions, like updating firewall rules or patching vulnerabilities, to strategic actions, such as user awareness campaigns or policy changes. In this section, we outline how AI can be harnessed to provide automated mitigation recommendations based on analyzed intelligence, and to visualize the integrated approach of AI-driven mitigation recommendations, a flowchart is presented in Figure 6.

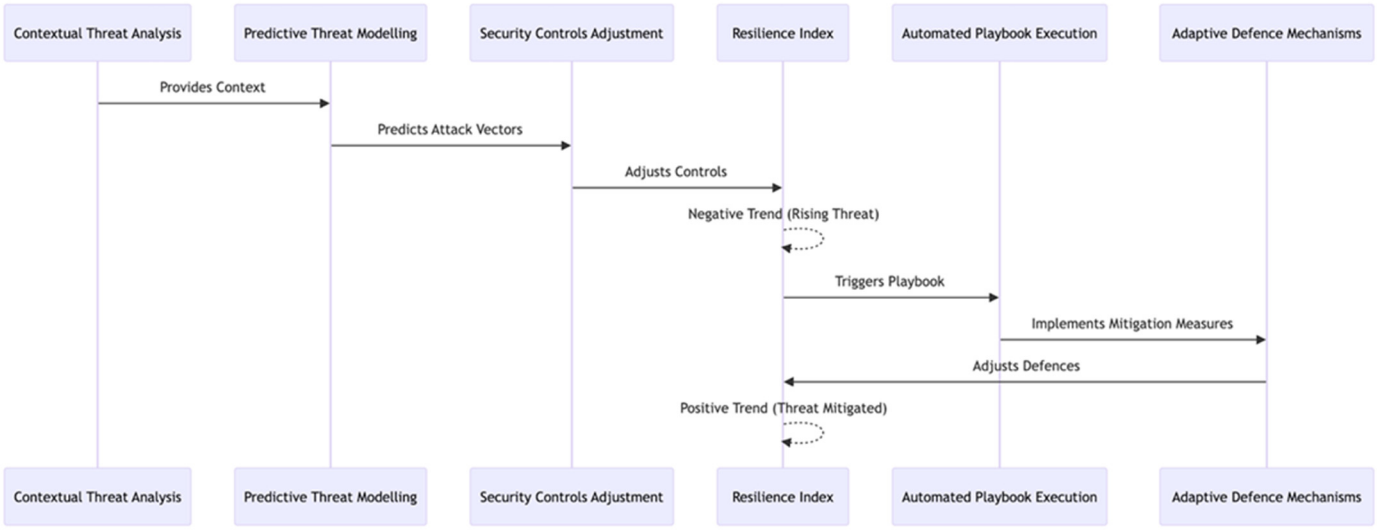

Figure 6. AI-driven CTI pipeline security control steering.

3.3.1. Contextual Threat Analysis

Before the AI-enhanced CTI pipeline recommends any mitigation strategies, it is crucial to understand the context of the threat against the operating IT landscape. Therefore, analysis of the threat in relation to the organization's infrastructure, assets, and previous incidents is necessary. This is achieved using natural language processing (NLP) to extract contextual information from threat intelligence reports [46].

Suppose an organization receives a threat intelligence report about a new ransomware strain targeting financial institutions. Using NLP, the AI system extracts keywords like "ransomware," "financial institutions," and cross-references these with the organization's IT and security landscape to determine the relevance and potential impact, thereby adjusting the relevant security controls accordingly.

3.3.2. Predictive Threat Modelling

AI trained with current intelligence and historical data can predict the likely progression of a threat based on patterns observed in past incidents [47]. As a result, we can preemptively strengthen defenses in vulnerable areas. For example, if historical data indicates that every time there is a spike in traffic from a particular region, a DDoS attack follows, the AI can predict a potential DDoS attack when it observes a similar traffic pattern in the future and, therefore, proactively adjust the relevant security controls accordingly.

3.3.3. Real-Time Threat Scoring

Not all threats have the same level of severity or relevance to an organization. AI provides a real-time threat score based on the specific IT landscape and organizational information, helping to prioritize mitigation efforts. As a result, it can score threats faster based on factors like the potential impact, exploitability, and the organization's vulnerability faster [48]. As a result, the most critical threats are addressed first. For instance, an organization might receive thousands of alerts daily. An AI system can score a detected phishing attempt as "high" risk if it is linked to a known APT group, while a generic malware detection might be scored as "medium" risk.

3.3.4. Automated Playbook Execution

For known threats or attack patterns, AI automatically executes predefined mitigation playbooks, reducing the response time subject to integration with threat intelligence systems and security orchestration, automation, and response (SOAR) platforms. Upon detecting a recognized threat pattern, the pipeline triggers the corresponding playbook, for immediate action. For example, if the AI pipeline detects patterns consistent with the "Emotet" malware, it can trigger a predefined playbook that isolates affected systems, blocks associated IPs, and sends notifications to the incident response team. Although the AI-enhanced CTI pipeline speeds up the threat response by running preset mitigation strategies, it is imperative to involve humans in key decision making to handle high-risk situations. For instance, cybersecurity experts should review and authorize actions chosen by AI in high-risk scenarios, combining the speed of AI with human insight to minimize the risk of automated responses.

3.3.5. Adaptive Defense Mechanisms

AI dynamically adjusts defenses based on ongoing threat analysis, using reinforcement learning (RL) models that can adapt security configurations in real-time [49]. For instance, if the AI pipeline detects increased traffic from a specific IP range associated with malicious activities, it can dynamically adjust the firewall rules to block or throttle that traffic. Or if the AI pipeline observes that every Friday evening there is an attempt to exfiltrate data, it can dynamically adjust the egress firewall rules during that time to add an additional layer of scrutiny.

3.3.6. Continuous Improvement

After implementing mitigation measures, it is essential to assess their effectiveness and refine the strategies accordingly. For instance, after blocking a suspected malicious IP address, the AI system can monitor for any subsequent attempts or changes in attack patterns from that IP address, refining the threat profile over time.

3.4. Resilience Verification

The simulation or emulation of attack scenarios, based on the received cyber threat intelligence, to assess the resilience of the implemented measures, is imperative at this stage. A security control effectiveness evaluation measures the resilience of organizations against potential cyber threats, eventually producing an alternative to a stock market index, but for cybersecurity. Much like financial indices, which provide traders with a snapshot of the market's health or trends, for e.g., the S&P 500 index, a cyber resilience index can offer decision makers a quick overview of their organization's cybersecurity posture. This index, updated by the AI in real-time, can serve as a barometer of an organization's cyber health and, therefore, can be used by decision makers to steer their defenses and resources accordingly. As a result, organizations are not just reactive, but proactive, in their cybersecurity approach.

To achieve this, we define the following three steps for resilience verification leading to the formation of a cyber resilience index: (1) automated penetration testing, (2) continuous monitoring, and (3) continuous improvement, as illustrated in Figure 7.

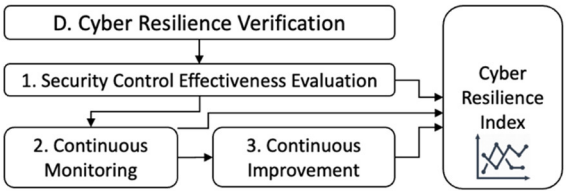

Figure 7. Cyber Resilience verification steps.

3.4.1. Security

AI provides a governance layer in regard to processes or tools to simulate or emulate cyber-attacks on a system, to identify vulnerabilities and assess the effectiveness of the implemented mitigation measures. As a result, AI simulates and emulates advanced persistent threats (APTs) to assess how well a system can withstand prolonged, targeted attacks. Moreover, since AI governs the CTI pipeline, it can adapt the strategies of APTs based on the received CTI and counter to the system's responses, mimicking the behavior of real-world adversaries [50]. Therefore, AI can provide a factual security control effectiveness validation rather than a checklist-based theoretical assessment.

3.4.2. Continuous Monitoring

AI continuously monitors the network traffic, system logs, and other relevant data sources throughout the IT landscape in which it is deployed. As a result, it provides real-time detection of suspicious activity that might indicate a breach or vulnerability exploitation. It can also be trained on historical network traffic data to recognize patterns related to known cyber threats, thus serving as an AI-powered intrusion detection system (IDS). Once deployed, it can monitor network traffic in real-time, flagging any deviations from the norm for further investigation.

3.4.3. Continuous Improvement

One of the key advantages of integrating AI into resilience verification is its ability to continuously self-improve. As an example, AI has the ability to observe and evaluate the

ecosystem in which it operates and can gain knowledge from any newly identified threats or weaknesses, improving its algorithms for the next evaluation. For instance, when an intrusion detection system identifies a new form of malware on the network, it can adjust the AI algorithms to identify this threat in the future. This continuous learning helps to keep the pipeline updated with the latest threat intelligence, coupled with the latest data from the IT landscape it is deployed on.

4. Challenges and Considerations

Although an AI-enhanced CTI processing pipeline may offer significant capabilities, it comes with several challenges and considerations that we discuss in this section. Given that we are in the nascent stages of building AI tools for corporate use, it is imperative for organizations starting to work with AI technologies, and especially within CTI, that they should follow guidelines set by NIST AI RMF 1.0 [51], the EU AI Act [52,53], and ISO 5338 [54]. Adhering to these standards will help organizations develop AI systems that are effective, efficient, ethically responsible, and compliant with global regulations.

4.1. Ethical Consideration in AI-Enhanced CTI Processing

4.1.1. Data Privacy and Confidentiality

AI-enhanced CTI processing requires access to vast amounts of data, some of which may be sensitive or confidential. Such datasets must be managed ethically, with respect to privacy laws and regulations. Organizations, therefore, should consider anonymizing the data used for training AI models to ensure that personally identifiable information (PII) is sufficiently protected. Moreover, data breaches or misuse can lead to significant reputational damage and legal consequences. It is, therefore, essential to implement strict data anonymization techniques, using differential privacy, while ensuring that data are stored securely and encrypted [55].

4.1.2. Consent, Surveillance, and Proportionality

While hunting for threat vectors, evidence, and indicators of compromise, either manually or with the use of AI, oftentimes the lines between legitimate surveillance and invasion of privacy are blurred. Hence, any data collection or surveillance activities should be executed with the necessary prior consent and in accordance with legal and ethical standards. Ultimately, the relevant individuals should be aware of and agree to the monitoring and data collection processes, respecting their autonomy and rights. Nonetheless, the implementation of clear consent mechanisms, regularly updating terms of service, and transparency in regard to data collection practices can solve this challenge [56]. The scale of the surveillance must also be proportionate, thus preventing overreach and the potential misuse of data. Solving this challenge requires regular audits and setting clear boundaries on data collection [57].

4.1.3. Technology Misuse

Like any innovative tool, the AI-enhanced CTI pipeline can potentially be misused. There is a potential risk of being used for malicious purposes, for instance, spreading misinformation or running unauthorized surveillance. Organizations should implement strict access controls and behavioral monitoring to prevent misuse. NIST recently provided a thorough AI risk management framework describing this challenge and potential solutions [51].

4.2. Addressing Potential Bias in AI-Enhanced CTI Models

4.2.1. Training Data Scrutiny

AI models are only as good as the data they are trained on. If the training data contains biases, the AI model will likely inherit those biases. It is crucial to use training datasets that are diverse and representative to avoid unintentional bias in AI-driven insights.

Biased training data will lead to skewed AI predictions, ultimately leading to unfair outcomes. To address this challenge, we need to consider diverse datasets, employing fairness-enhancing interventions, and regular bias audits. However, even with unbiased training data, algorithms themselves can introduce biases [58]. It is, therefore, evident that regular audits and evaluations of AI models can help to identify and rectify any inherent biases in AI algorithms [59].

4.2.2. Continuous Model Evaluation

Continuous model evaluation will provide reasonable assurance that the AI-enabled pipeline will remain relevant and unbiased as new CTI data emerges. Moreover, biases in the AI model may create loops where the model's predictions reinforce existing biases. For instance, if an AI model incorrectly flags certain types of network traffic as malicious due to bias, it may lead to increased scrutiny of similar traffic in the future, reinforcing the bias. It is, therefore, imperative to implement real-time evaluation metrics and periodic retraining of the models [59]. If the AI-driven CTI pipeline makes a wrong decision, it is crucial to have a reliable rollback mechanism in place to restore the system to its previous working state. This can be done by creating epochs or checkpoints before carrying out any playbook actions [60]. Therefore, if a decision is found to be incorrect or harmful, the system can easily go back to a state before the action was taken, reducing the likelihood of disruptions.

4.2.3. Systematic Bias Detection

The AI-enhanced CTI pipeline may be prone to observer bias, which appears when the subjective predispositions of individuals involved in the AI's development or operational phases influence the selection of training data or the interpretation of the system's outputs. Such biases can inadvertently lead to misrepresentation in the AI model, both in the CTI, as well as in regard to its threat detection capabilities. This will potentially result in the disproportionate identification or neglect of specific threats. To protect the precision and impartiality of the AI-enhanced CTI pipeline, it is imperative to, firstly, acknowledge and address the presence of observer bias. Moreover, training data bias, a form of availability bias, originates from the initial CTI collection phase of the pipeline. This may occur when the training data predominantly consists of easily accessible information, or when over or under sampling leads to a training dataset that does not accurately represent real-world scenarios. Nonetheless, mitigations for both these biases have been proposed by [61] and other scholars [62] such as the utilization of heterogeneous data sources, and the implementation of systematic bias detection and correction mechanisms throughout the lifecycle of the AI model. Additionally, organizations should consider regularly testing the AI system's performance using blind or double-blind methods, where neither the testers nor the AI system has information that might influence the outcome of the test. This would provide a method of assessing the AI system's ability to identify threats without bias.

4.3. Transparency and Interpretability within the AI-Enhanced CTI Pipeline

4.3.1. Explainable AI (XAI) and Stakeholder Trust

Many advanced AI models, especially deep learning models, are oftentimes seen as "black boxes", where their decision-making processes are not easily interpretable. This poses a significant challenge in regard to CTI, where understanding the rationale behind insights is crucial for decision making. XAI enables AI model predictions to become understandable by humans, fostering trust, and facilitating better decision making. To address the black box issue, there is a growing amount of stress on model explainability. SHAP (SHapley Additive exPlanation) or LIME (Local Interpretable Model-Agnostic Explanation) techniques have emerged as a means to provide insights into how AI models arrive at their decisions. Therefore, using interpretable models, employing post-hoc

explanation techniques, and visualizing model decisions are demonstrated ways forward to solve this challenge effectively [63].

Lastly, it is of utmost importance for AI-driven CTI systems to be effective, that stakeholders must trust the insights with which they are provided. Guaranteeing transparency and interpretability is key to building and maintaining this trust. Regular communication about how the AI models work, their limitations, and the steps taken to address inaccuracy can help foster trust among stakeholders.

4.3.2. Feedback within the Human—AI fusion

Feedback from CTI analysts can refine AI models and, therefore, verify their continued relevance and accuracy. This can be achieved through the implementation of an iterative refinement model [64], and by fostering a collaborative AI–human environment. For instance, a CTI analyst provides real-world insights and corrections, which can be used to fine tune AI algorithms. Or, if an AI model misclassifies a type of malware, feedback from analysts can correct this mistake, leading to improved future detection capabilities. Additionally, comprehensive documentation warrants that all stakeholders understand the workings, limitations, and scope of AI models, thus maintaining detailed model logs, providing clear documentation on algorithms and training data, and confirming transparency in model updates [65].are all important.

4.4. AI Model Robustness and Adversarial Attacks

Prior to adopting an AI-driven CTI pipeline, a prime consideration is that the pipeline itself might become a potential target for cyber adversaries. Similar to traditional software systems that can be exploited, AI models have their own set of vulnerabilities, especially to adversarial attacks. These attacks involve feeding the model specially crafted input data designed to deceive it, leading to incorrect outputs or predictions [66]. Adversarial attacks can be categorized on a high level, into two types. White-box attacks, where the attacker has complete knowledge of the AI model, the architecture, and its weights, and black-box attacks, where the attacker has no knowledge of the model's internal aspects and only has access to its inputs and outputs. Common adversarial attack techniques include adding imperceptible noise to input data, generating adversarial examples, or exploiting model transferability, where an adversarial example crafted for one model affects another [67]. An adversarial attack could lead to several adverse outcomes, for instance, the misclassification of benign network traffic as malicious, or vice versa, incorrect threat scoring, leading to missed prioritization of threats, or even deceptive insights that could mislead incident response teams or decision-makers [67].

To protect the AI-driven CTI pipeline against adversarial attacks, several strategies can be followed. For instance, training the AI model on adversarial examples, making it more robust to such attacks. Input filtering or normalization can detect and mitigate adversarial input data [68]. The use of collaborative models will increase its robustness, as an attacker would need to deceive multiple models simultaneously. Moreover, continuously updating and retraining the AI model confirms that it is equipped to manage new adversarial techniques. Organizations should also consider adopting frameworks to perform red teaming against generative AI models, such as PyRIT v0.2.1 (https://www.microsoft.com/en-us/security/blog/2024/02/22/announcing-microsofts-open-automation-framework-to-red-team-generative-ai-systems/, accessed on 14 January 2024), to assess and improve the security posture of AI models, making them more resilient to adversarial threats and countering bias within the models. Lastly, implementing real-time monitoring to detect unusual patterns in the model's predictions can flag potential adversarial attacks.

5. Conclusions and Future Research

In this work, we defined an AI-enhanced CTI processing pipeline and presented how the integration of AI into CTI processing has the potential to revolutionize the cybersecurity landscape. AI can automate, enhance, and expedite numerous CTI tasks, starting with augmenting the human analyst in separating the "signal" from the "noise", namely actionable cyber intelligence. Moving on to data ingestion, collaborative analysis, automated mitigation and, lastly, to cyber resilience verification. Therefore, organizations can achieve a more proactive and adaptive cybersecurity posture, staying a step ahead of the evolving threats, as opposed to adopting a reactive approach. Successful implementation would signal the beginning of an AI-based end-to-end cyber defense system. Currently, human CTI analysts and AI are connected by strong collaborative bonds, where both complement each other's strengths and counter each other's weaknesses.

However, the integration of AI into CTI brings implementation challenges, ethical considerations, potential biases, and the need for transparency and interpretability that require attention. Nonetheless, with a balanced approach that combines the strengths of AI with human expertise, these challenges can be addressed effectively.

With this work, we set the foundational framework for an AI-enhanced CTI processing pipeline, and we identify several research routes. Some potential directions for future research may involve advanced AI models. As AI continues to evolve, exploring newer models and architectures tailored for CTI tasks could yield even more accurate and efficient results. Another angle is the ethical use of AI in cybersecurity. Thorough research into the ethical implications of AI use in cybersecurity would be useful, alongside the development of guidelines and best practices for responsible deployment.

Moreover, on the Human–AI collaboration aspect, further research on optimizing the collaboration between AI systems and human analysts could demonstrate how each complements the other's strength. Researching the integration of more diverse and unconventional data sources into the CTI pipeline, such as social media, could potentially allow for a more accurate prediction of the threat actor's next attacks. Another research route is to study ways to mitigate bias in the pipeline and explore ways to make risk decisions about which runbacks can be automated and which cannot. Lastly, a more technical angle would be to investigate the feasibility and methodologies for real-time threat intelligence processing using AI, enabling an instantaneous response to emerging threats, while also achieving automated compliance with security policies, which is our next focus.

Author Contributions: Conceptualization, L.A.; Investigation, L.A; Writing—original draft, L.A.; Writing—review & editing, M.D; Supervision, M.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Johnson, C.; Badger, L.; Waltermire, D.; Snyder, J.; Skorupka, C. National Institute of Standards and Technology—U.S Department of Commerce. October 2016. Available online: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-150.pdf (accessed on 22 September 2023).
2. Phythian, M. Studies in Intelligence. In *Understanding the Intelligence Cycle*; Routledge Taylor & Francis Group: London, UK; New York, NY, USA, 2013; pp. 21–43.
3. Buczak, A.; Guven, E. A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Commun. Surv.* **2016**, *18*, 1153–1176.
4. Strom, B.E.; Applebaum, A.; Miller, D.P.; Nickels, K.C.; Pennington, A.G.; Thomas, C.B. The MITRE Corporation. March 2020. Available online: https://attack.mitre.org/docs/ATTACK_Design_and_Philosophy_March_2020.pdf (accessed on 27 September 2023).
5. Dekker, M.; Alevizos, L. A threat-intelligence driven methodology to incorporate uncertainty in cyber risk analysis and enhance decision-making. *Wiley Secur. Priv.* **2023**, *7*, e333.

6. Chen, T.M.; Abu-Nimeh, S. Lessons from Stuxnet. *Computer* **2011**, *44*, 91–93.
7. Sarker, I.H.; Furhad, H.M.; Nowrozy, R. AI-Driven Cybersecurity: An Overview, Security Intelligence Modeling and Research Directions. *SN Comput. Sci.* **2021**, *2*, 173.
8. Ring, M.; Wunderlich, S.; Grudl, D. Flow-based benchmark data sets for intrusion detection. In Proceedings of the 16th European Conference on Cyber Warfare and Security, Dublin, Ireland, 29–30 June 2017; p. 361.
9. Brown, R.; Nickels, K. *SANS 2023 CTI Survey: Keeping Up with a Changing Threat Landscape*; SANS Institute: Boston, MA, USA, 2023.
10. Sauerwein, C.; Sillaber, C.; Mussmann, A.; Breu, R. Threat Intelligence Sharing Platforms: An Exploratory Study of Software Vendors and Research Perspectives. In Proceedings of the der 13 Internationalen Tagung Wirtschaftsinformatik (WI 2017), St. Gallen, Switzerland, 12–15 February 2017.
11. Sundar, S.S. Rise of Machine Agency: A Framework for Studying the Psychology of Human–AI Interaction (HAII). *J. Comput. - Mediat. Commun.* **2020**, *25*, 74–88.
12. Brundage, M.; Avin, S.; Wang, J.; Belfield, H.; Krueger, G.; Hadfield, G.; Khlaaf, H. arXiv—Computer Science—Computers and Society. 20 April 2020. Available online: https://arxiv.org/abs/2004.07213 (accessed on 30 September 2023).
13. Varma, A.J.; Taleb, N.; Said, R.A.; Ghazal, T.M.; Ahmad, M.; Alzoubi, H.M.; Alshurideh, M. A Roadmap for SMEs to Adopt an AI Based Cyber Threat Intelligence. In *The Effect of Information Technology on Business and Marketing Intelligence Systems*; Springer: Cham, Switzerland, 2023; pp. 1903–1926.
14. Suryotrisongko, H.; Musashi, Y.; Tsuneda, A.; Sugitani, K. Robust Botnet DGA Detection: Blending XAI and OSINT for Cyber Threat Intelligence Sharing. *IEEE Access* **2022**, *10*, 34613–34624.
15. Ranade, P.; Piplai, A.; Mittal, S.; Joshi, A.; Finin, T. arXiv—Computer Science—Cryptography and Security. 18 June 2021. Available online: https://arxiv.org/abs/2102.04351 (accessed on 1 October 2023).
16. Moraliyage, H.; Sumanasena, V.; De Silva, D.; Nawaratne, R.; Sun, L.; Alahakoon, D. Multimodal Classification of Onion Services for Proactive Cyber Threat Intelligence Using Explainable Deep Learning. *IEEE Access* **2022**, *10*, 56044–56056.
17. Mitra, S.; Piplai, A.; Mittal, S.; Joshi, A. Combating Fake Cyber Threat Intelligence using Provenance in Cybersecurity Knowledge Graphs. In Proceedings of the IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 15–18 December 2021.
18. Mittal, S.; Joshi, A.; Finin, T. arXiv. 7 May 2019. Available online: https://arxiv.org/pdf/1905.02895.pdf (accessed on 3 October 2023).
19. Kotsiantis, S. Decision trees: A recent overview. *Artif. Intell. Rev.* **2011**, *39*, 261–283.
20. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; The MIT Press: London, UK, 2016.
21. Nguyen, H.; Tran, K.; Thomassey, S.; Hamad, M. Forecasting and Anomaly Detection approaches using LSTM and LSTM Autoencoder techniques with the applications in supply chain management. *Int. J. Inf. Manag.* **2021**, *57*, 102282.
22. Landwehr, N.; Hall, M.; Frank, E. Logistic Model Trees. *Mach Learn* **2005**, *59*, 161–205.
23. Chen, H.; Chiang, R.H.L.; Storey, V.C. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Q.* **2012**, *36*, 1165–1188.
24. Tian, K.; Revelle, M.; Poshyvanyk, D. Using Latent Dirichlet Allocation for automatic categorization of software. In Proceedings of the 6th IEEE International Working Conference on Mining Software Repositories, Vancouver, BC, Canada, 16–17 May 2009.
25. Schonlau, M.; Zou, R.Y. The random forest algorithm for statistical learning. *Stata J. Promot. Commun. Stat. Stata* **2024**, *20*, 3–29.
26. Sarker, I.H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Comput. Sci.* **2021**, *2*, 1–20.
27. Marschall, O.; Cho, K.; Savin, C. A Unified Framework of Online Learning Algorithms for Training Recurrent Neural Networks. *J. Mach. Learn. Res.* **2020**, *21*, 1–34.
28. Catak, F.O.; Kuzlu, M.; Catak, E.; Cali, U.; Guler, O. Defensive Distillation-Based Adversarial Attack Mitigation Method for Channel Estimation Using Deep Learning Models in Next-Generation Wireless Networks. *IEEE Access* **2022**, *10*, 98191–98203.
29. Settles, B. Synthesis Lectures on Artificial Intelligence and Machine Learning (SLAIML). In *Active Learning*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 1–114.
30. Kuhl, N.; Goutier, M.; Baier, L.; Wolf, C.; Martin, D. Human vs. supervised machine learning: Who Learn. Patterns Faster? *Cogn. Syst. Res.* **2022**, *76*, 78–92.
31. Mishra, S. An Optimized Gradient Boost Decision Tree Using Enhanced African Buffalo Optimization Method for Cyber Security Intrusion Detection. *Appl. Sci.* **2023**, *12*, 12591.
32. Desmond, M.; Muller, M.; Ashktorab, Z.; Dugan, C.; Duesterwald, E.; Brimijoin, K.; Finegan-Dollak, C.; Brachman, M.; Sharma, A. Increasing the Speed and Accuracy of Data Labeling Through an AI Assisted Interface. In Proceedings of the IUI '21: 26th International Conference on Intelligent User Interfaces, College Station, TX, USA, 13–17 April 2021.
33. Chen, C.P. Deep learning for pattern learning and recognition. In Proceedings of the IEEE 10th Jubilee International Symposium on Applied Computational Intelligence and Informatics, Timisoara, Romania, 21–23 May 2015.
34. Angelelli, M.; Arima, S.; Catalano, C.; Ciavolino, E. arXiv. 1 August 2023. Available online: https://arxiv.org/abs/2302.08348 (accessed on 12 April 2023).
35. Lemay, A. Leblanc and Sylvain. Cognitive Biases in Cyber Decision-Making. In Proceedings of the ICCWS 2018 13th International Conference on Cyber Warfare and Security, Washington, DC, USA, 8–9 March 2018.
36. Kartal, E. A Comprehensive Study on Bias in Artificial Intelligence Systems: Biased or Unbiased AI, That's the Question! *Int. J. Intell. Inf. Technol. (IJIIT)* **2022**, *18*, 1–23.
37. Lorente, A. Setting the goals for ethical, unbiased, and fair AI. . In *AI Assurance*; Academic Press: Cambridge, MA, USA, 2023; pp. 13–54.

38. Dell'Acqua, F.; McFowland, E., III; Mollick, E.; Lifshitz-Assaf, H.; Kellogg, K.C.; Rajendran, S.; Krayer, L.; Candelon, F.; Lakhani, K.R. Harvard Business School. 22 September 2024. Available online: https://www.hbs.edu/ris/Publication%20Files/24-013_d9b45b68-9e74-42d6-a1c6-c72fb70c7282.pdf (accessed on 2 February 2024).
39. Kaloudi, N.; Li, J. The AI-Based Cyber Threat Landscape: A Survey. *ACM Comput. Surv.* **2020**, *53*, 1–34.
40. Sarker, I.H. AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems. *SN Comput. Sci.* **2022**, *3*, 1–20.
41. Gupta, C.; Johri, I.; Srinivasan, K.; Hu, Y.-C.; Qaisar, S.M.; Huang, K.-Y. A Systematic Review on Machine Learning and Deep Learning Models for Electronic Information Security in Mobile Networks. *Sensors* **2022**, *22*, 2017.
42. Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput. Sci.* **2021**, *2*, 160.
43. Djenna, A.; Bouridane, A.; Rubab, S.; Marou, I.M. Artificial Intelligence-Based Malware Detection, Analysis, and Mitigation. *Symmetry* **2023**, *15*, 677.
44. Akhtar, M.S.; Feng, T. Malware Analysis and Detection Using Machine Learning Algorithms. *Symmetry* **2022**, *14*, 2304.
45. Mohamed, N. Current trends in AI and ML for cybersecurity: A state-of-the-art survey. *Cogent Eng.* **2023**, *10*, 2.
46. Jain, J. Artificial Intelligence in the Cyber Security Environment. In *Artificial Intelligence and Data Mining Approaches in Security Frameworks*; Wiley: Hoboken, NJ, USA, 2021.
47. Sree, S.V.; Koganti, S.C.; Kalyana, S.K.; Anudeep, P. Artificial Intelligence Based Predictive Threat Hunting in the Field of Cyber Security. In Proceedings of the 2nd Global Conference for Advancement in Technology (GCAT), Bangalore, India, 1–3 October 2021.
48. Gupta, I.; Gupta, R.; Singh, A.K.; Wen, X. An AI-Driven VM Threat Prediction Model for Multi-Risks Analysis-Based Cloud Cybersecurity. *Trans. Syst. Man Cybern. Syst.* **2023**, *53*, 6815–6827.
49. Deep Reinforcement Learning for Cyber Security. *Trans. Neural Netw. Learn. Syst.* **2023**, *34*, 3779–3795.
50. Confido, A.; Ntagiou, E.V.; Wallum, M. Reinforcing Penetration Testing Using AI. In Proceedings of the 2022 IEEE Aerospace Conference (AERO), Big Sky, MT, USA, 5–12 March 2022.
51. NIST. National Institute of Standards and Technology—U.S. Department of Commerce. January 2023. Available online: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf (accessed on 8 October 2023).
52. P. R. Committee. Council of the European Union. 25 November 2022. Available online: https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf (accessed on 12 January 2024).
53. Council of the European Union. 26 January 2024. Available online: https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf (accessed on 4 February 2024).
54. ISO. International Standards Organization. 2023. Available online: https://www.iso.org/standard/81118.html (accessed on 18 January 2024).
55. Sweeney, L. k-Anonymity: A Model for Protecting Privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **2002**, *10*, 557–570.
56. Solove, D.J. Privacy Self-Management and the Consent Dilemma. *Harv. Law Rev.* **2013**, *126*, 1880.
57. Tene, O.; Polonetsky, J. Big Data for All: Privacy and User Control in the Age of Analytics. *J. Technol. Intellect. Prop.* **2013**, *11*, 240–272.
58. Solon, B.; Selbst, D.A. Big Data's Disparate Impact. *Calif. Law Rev.* **2016**, *104*, 671–732.
59. Danks, D.; London, A.J. Algorithmic Bias in Autonomous Systems. In Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI 2017), Pittsburgh, PA, USA, 19–25 August 2017.
60. Abdullah, I.U.T. arXiv. 27 May 2023. Available online: https://arxiv.org/pdf/2305.19298.pdf (accessed on 6 October 2023).
61. Schwartz, R.; Vassilev, A.; Greene, K.K.; Perine, L. National Institute of Standards and Technology (NIST). 15 March 2022. Available online: https://doi.org/10.6028/NIST.SP.1270 (accessed on 24 February 2024).
62. Ha, T.; Kim, S. Improving Trust in AI with Mitigating Confirmation Bias: Effects of Explanation Type and Debiasing Strategy for Decision-Making with Explainable AI. *Int. J. Hum. –Comput. Interact.* **2023**, 1–12.
63. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the KDD '16: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016.
64. Holstein, K.; Vaughan, J.W.; Daume, H.; Dudik, M.; Wallach, H. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? In Proceedings of the CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019.
65. Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J.W.; Wallach, H.; Daumé, H.; Crawford, K. arXiv—Computer Science—Databases. 1 December 2021. Available online: https://arxiv.org/abs/1803.09010 (accessed on 14 October 2023).
66. Vassilev, A.; Oprea, A.; Fordyce, A.; Anderson, H. National Institute of Standards and Technology. January 2024. Available online: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf (accessed on 25 January 2024).
67. Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain. *ACM Comput. Surv.* **2021**, *54*, 5.
68. Kaur, R.; Gabrijelčič, D.; Klobučar, T. Artificial intelligence for cybersecurity: Literature review and future research directions. *Inf. Fusion* **2023**, *97*, 101804.