False Sense of Security Caused by AI-Driven Threat Detection

Parth Talaviya[1], Manya Parikh2, Deep Solanki3, and Jigar Gajjar4

Lok Jagruti Kendra University, Ahmedabad, India[1]
Lok Jagruti Kendra University, Ahmedabad, India[2]
TechD Cybersecurity Limited, Ahmedabad, India[3]
TechD Cybersecurity Limited, Ahmedabad, India[4]

talaviyaparth604@gmail.com[1],manyaparikh23@gmail.com[2],deepsolanki5799@gmail.com[3],cyberian.jg@gmail.com[4]

Abstract: Artificial Intelligence (AI) has quickly proven its relevance in the current state of the cybersecurity environment, primarily because of its ability to help organizations identify possible threats more quickly and efficiently. By constantly analyzing large amounts of data related to the security environment, AI-powered detection systems can detect unusual patterns of behavior that could potentially go unnoticed. This has greatly improved the monitoring process and the early identification of threats. However, the increasing dependence on automated detection systems also poses a possible threat, as organizations become more reliant on these systems and may feel more secure than they actually are. This paper will discuss the risks posed by these dependencies through the analysis of the current literature available in the academic community, as well as through a great deal of experience in the cybersecurity environment. It will cover topics such as automation bias, adversarial attacks, and the probabilistic nature of algorithmic decision-making, all of which pose possible risks to the current state of the security environment. To counter these risks, this paper will propose a Controlled Trust Scheme that aims to strike a balance between the efficiency of machine-based systems and human oversight. The findings show that the effectiveness of the current cybersecurity environment not only depends on technological development but also on proper governance to ensure that the confidence of the organization is in line with its actual preparedness.

Keywords: Artificial Intelligence, Cyber Threat Detection, Automation Bias, Security Analytics, Machine Learning Security, Cyber Risk Management, Intelligent Defense

I. INTRODUCTION

The rate at which the digital world is evolving in terms of business, infrastructure, and communication has led to a certain level of connectivity that would have been difficult to predict a decade ago. While this interconnected system of systems has led to a huge increase in efficiency and innovation, it has also led to an increase in the attack surface that is exposed to cyber attackers. Cybersecurity professionals are no longer tasked with securing systems that are isolated; rather, they are tasked with securing highly dynamic environments where data is constantly flowing from devices, cloud systems, and remote access points.

In the early stages of the evolution of the cybersecurity industry, threat protection was largely dependent on signature-based and rule-based systems. These systems were sufficient while attack behavior was predictable and followed known patterns. However, as cyber attackers began to employ more dynamic attack behavior, the limitations of static threat protection systems became more and more apparent. Attack behavior outpaced the development of signature databases, and it became apparent that more dynamic defensive technologies were necessary to effectively mitigate an increasingly unpredictable threat environment.

However, Artificial Intelligence soon emerged as a promising approach to overcome these challenges. Rather than focusing on known malicious code, AI-powered approaches focus on activity patterns to detect anomalies that might possibly indicate an impending attack. This enables organizations to detect malicious activity even in the face of novel attack patterns [1], [2]. However, technological development can also affect organizational culture. With the success of AI-powered approaches being established, there is a rising trend to fall back on the assumption that technological development per se provides comprehensive security coverage. This is a very perilous trend, as the success of cybersecurity solutions is not contingent on technological development but also on the prudent use of technology.

http://doi.org/10.63766/spujstmr.26.000085

In an effort to acquire a comprehensive understanding of this complex phenomenon, it is imperative to examine digital threat detection and its significance in modern-day cybersecurity.

A. What is Digital Threat Detection and Why is it Important in Modern Cybersecurity?

Digital threat detection is the systematic approach to detecting potential actions that could impact the confidentiality, integrity, or availability of information systems. Instead of reacting to harm after it has occurred, detection tools are designed to detect early warning signs in time to contain the damage.

Cyber attacks are not typically immediate in their destructive capabilities. Most attacks will take a gradual approach, beginning with reconnaissance before moving on to further system exploitation. If left undetected, early warning signs could easily be masked among normal system traffic, allowing attackers to go undetected for an extended period of time. The value of detection has grown in proportion to the amount of data being produced within the operational environment. Human monitoring is insufficient to interpret millions of events on a daily basis, making it necessary to have automated support to maintain situational awareness [3]. Numerous studies have demonstrated that detection time has a direct correlation to lower breach costs and shorter recovery cycles [4].

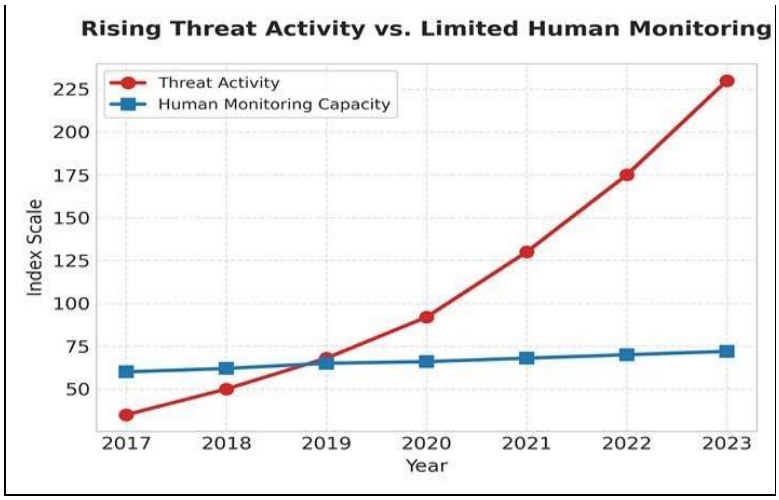

Fig.1. Expansion of Cyber Threat Activity vs Monitoring Capability

B. What is AI-Driven Threat Detection? Key Features and Capabilities

AI-powered threat detection can be defined as the use of machine learning algorithms and intelligent analytics to detect malicious activity in digital infrastructure. Unlike traditional security solutions that are mainly dependent on signature-based detection, AI-powered threat detection uses behavioral patterns and context clues to detect malicious activity. This enables organizations to detect unknown threats that would not be detectable by traditional security solutions.

The first major benefit of AI-powered threat detection is its ability to process a huge amount of security data in real-time. Security data from networks, authentication infrastructure, and enterprise applications can be processed simultaneously, which enables improved situational awareness and reduces the workload of security analysts [5]. Moreover, adaptive learning allows these solutions to optimize detection variables over time, which enables improved responsiveness to new threat tactics [6]. The main use cases of AI-powered threat detection include detecting anomalous activity, producing prioritized alerts, determining risk levels, and facilitating malware analysis. Although these use cases improve efficiency, it is important to recognize that algorithmic outputs are necessarily probabilistic. Hence, intelligent detection can be considered a decision-enabling technology and not a definitive source of security authority.

TABLE I

PRIMARY FUNCTIONS OF AI IN THREAT DETECTION

Function	Practical Role	Security Outcome
Behavioral Analysis	Identifies unusual activity	Earlier threat discovery
Automated Alerting	Flags high-risk events	Faster response cycles
Predictive Assessment	Estimates potential attacks	Preventive posture
Pattern Recognition	Detects malware similarities	Improved endpoint defense
Continuous Processing	Monitors environments 24/7	Reduced operational gaps

C. How AI is Integrated into Threat Detection Systems

The integration of AI in the cybersecurity framework has been gradual rather than sudden. Most organizations began by using machine learning as a supporting analytical component before integrating it into the detection process.Security Information and Event Management systems have incorporated intelligent engines that can correlate alerts from different sources. This allows analysts to determine whether isolated events are indicative of a larger threat pattern [7].Cloud computing has been the driving force behind this shift. Organizations now have the ability to process complex data without having to invest in expensive infrastructure. Another area of interest is the emergence of User and Entity Behavior Analytics, where systems set up behavioral profiles and alert on anomalies that could be indicative of credential attacks or insider threats. Notwithstanding the benefits, there is a subtle dependency that arises when analysis by machines becomes the norm, and challenging system output becomes less common.

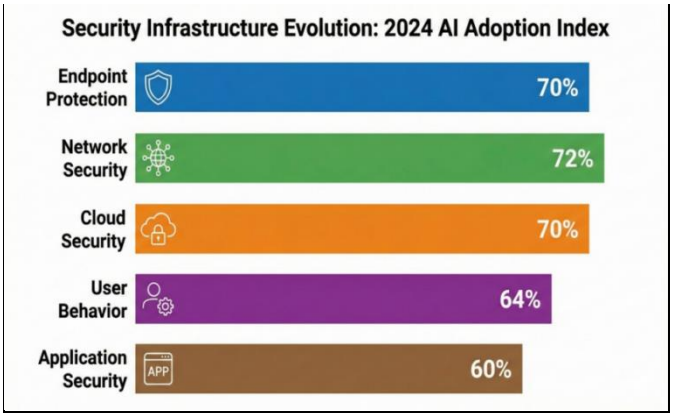

Fig.2.Distribution of AI Usage Across Security Domains

TABLE II

TRADITIONAL DETECTION VS AI-ENABLED DETECTION

Evaluation Area	Traditional Approach	AI-Enabled Approach
Detection Method	Signature matching	Behavioral modeling
Adaptability	Requires manual updates	Learns from data
Investigation Speed	Analyst-dependent	Near real-time
Unknown Threat Recognition	Limited	Stronger capability
Operational Load	High	Moderately reduced

D. Advantages of AI-Based Threat Detection Over Traditional Security Methods

The shift toward intelligent detection has been driven largely by operational necessity. Security teams face an environment where the volume of alerts can exceed human processing capacity, making automation a practical requirement rather than a technological luxury.AI contributes significantly by filtering irrelevant signals and directing attention toward events that merit investigation. This prioritization helps reduce analyst fatigue, a factor frequently associated with missed warnings.Another advantage lies in contextual evaluation. Instead of examining events independently, AI systems can recognize relationships between behaviors — for example, linking unusual login activity with abnormal data transfers [8]. Cost efficiency also influences adoption decisions. While implementing intelligent platforms requires investment, automated workflows often reduce long-term staffing pressures.Still, technological advancement should not be mistaken for immunity. Every defensive mechanism operates within constraints.

E. The False Sense of Security: Risks of Overreliance on AI-Driven Detection

The more organizations become comfortable with automated monitoring, the more confidence in algorithmic security may build faster than critical assessment. With seemingly stable dashboards and a reduction in alert volumes, it is easy to see this calm as a sign of excellent defense. Human psychology is also at play. Automation bias, the tendency to rely on computer-suggested solutions, may demotivate analysts to question system decisions [9]. Attackers understand this phenomenon. Some attacks are designed to stay below anomaly thresholds, allowing persistence without immediate alerting [10]. It is not the use of AI that poses a threat but the potential for trust in technology to build faster than reality. It is vital to understand this difference in order to develop robust security plans.

II. LITERATURE SURVEY

A. Evolution of Cyber Threat Detection Systems

The development of cyber threat detection has been underway for a number of generations of technology, which have been driven by the dynamic nature of cyber threats. In the early days of cyber threat protection, the most common approaches to detection were signature-based software that scanned for new files against a database of known threats. While these approaches were very effective at detecting known threats, they were ineffective at detecting new and rapidly evolving threats. In an effort to make these approaches more effective, heuristic analysis was developed, which enabled security software to detect threats based on behavior rather than known attack patterns. Unfortunately, heuristic analysis was also prone to high levels of false positives, which further taxed security analysts.

he increasing complexity of enterprise networks eventually reached a point where manual analysis was no longer possible, and researchers began to investigate data-driven models for detection. Machine learning algorithms provided a possible solution to this problem, which enabled systems to detect anomalies within known behavior patterns rather than known attack patterns [11], [12]. The current state of detection represents this shift from static recognition to adaptive intelligence, which provides the basis for AI-based cybersecurity solutions.

B. Adoption of AI in Threat Detection and Security Operations

The increasing complexity of the cyber threat environment has led to the increasing adoption of Artificial Intelligence in the present security framework. The present framework produces a tremendous amount of telemetry data, ranging from network traffic to endpoint activity, making it difficult to manually analyze. Machine learning algorithms are used to overcome this difficulty by detecting behavioral anomalies and contextual indicators of potential threats, thus improving early-stage threat detection [13].

Breakthroughs in deep learning have furthered these developments by allowing the detection of structural relationships between malware and signature-based threats that are designed to evade traditional signature-based security frameworks [14]. On the implementation front, AI is now fully integrated with Security Information and Event Management (SIEM) systems, where correlation engines combine alerts from various sources to improve visibility in a distributed environment [15]. On the other hand, User and Entity Behavior Analytics (UEBA) is also employed for insider threat detection by creating behavioral profiles and pointing out anomalies such as unusual login attempts or unauthorized access to information [16].

However, despite these improvements, it has been warned that threat detection by AI is not completely free from challenges. The efficiency of AI algorithms is also dependent on the quality of the training data, and detection efficiency can be affected by an unfamiliar environment of operation [17]. Furthermore, there has been a rising trend among attackers to develop evasion techniques that may potentially be used to deceive detection algorithms, and this is the never-ending cycle of intelligent defense and adaptive attack tactics [18].

TABLE III

TRADITIONAL THREAT DETECTION VS AI-DRIVEN DETECTION

Dimension	Traditional Detection	AI-Driven Detection
Analytical Basis	Known signatures	Behavioral modeling
Adaptation Speed	Slow	Continuous learning
Alert Handling	Mostly manual	Automated prioritization
Detection Scope	Known threats	Known + emerging threats
Infrastructure Fit	On-premise focused	Cloud-compatible

TABLE IV

AI CAPABILITIES VS ASSOCIATED SECURITY RISKS

AI Capability	Security Benefit	Associated Risk
Automated triage	Faster investigations	Reduced human scrutiny
Predictive analytics	Early warning signals	False confidence
Behavioral detection	Identifies stealth activity	Context misinterpretation
Continuous monitoring	Minimal downtime	Silent failure possibility
Self-learning models	Adaptive defense	Model manipulation

C. Real-World Incidents Highlighting Overreliance on AI

Cybersecurity events in the real world demonstrate that even with intelligent detection systems, there are still risks involved. The Equifax hack is a case in point, where the attackers used an unpatched web application vulnerability and operated undetected for a long time. Although there were monitoring systems in place, operational control was ineffective in reacting to early warning signs, allowing a massive data exfiltration to take place. Analysis of studies on the breach timeline indicates that the delayed detection of the attack contributed significantly to its overall impact [19].

The SolarWinds supply chain attack is another case where attackers injected malicious code into software updates. Since the compromised traffic was legitimate, many automated security systems did not flag the activity as malicious immediately. It has been observed that attackers used trusted communication channels to evade monitoring layers, allowing them to operate stealthily in the target environments for a long time [20].

These examples highlight the following critical point: even with advanced detection systems, there may not be immediate recognition of the threat, especially when attackers target trust-based system behaviors.

D. Risks and Security Gaps Created by AI-Driven Threat Detection

Though AI has achieved remarkable progress in the field of defense, it is increasingly recognized that intelligent automation also poses vulnerabilities in addition to its benefits. One of the most widely recognized risks associated with automation is automation bias, which is defined as the human tendency to rely heavily on system outputs. As long as the volume of alerts appears to be under control, analysts often have a sense of operational security that can lead them to overlook the subtle warning signs [21].Another problem that the accuracy of detection poses is that most models are based on probabilistic domains. This implies that some malicious activities might remain entirely undetected. Attackers are also aware of this reality and often plan their attacks in a manner that is extremely similar to their benign versions, thereby reducing the possibility of detection. The

growing trend of adversarial machine learning also testifies to this reality, as it has been demonstrated that well-crafted inputs can influence the algorithmic interpretation of models and thereby impact detection accuracy [22].

The efficacy of the model in the long run is also affected by the phenomenon of model drift. As the organizational environment keeps changing, models that were accurate in the past may become less relevant unless they are constantly updated. The issue of interpretability also adds to this problem, as most AI systems tend to be relatively opaque, making it difficult for analysts to understand why a particular decision was made. When the output is not easily interpretable, critical assessment tends to fall by the wayside, thus inadvertently contributing to a false sense of security. There is also a potential for systemic vulnerability that arises from over-reliance on centralized analytics. If the detection pipelines are exposed to disruption or latency, organizations may temporarily be left in the dark about their security posture. Taken together, these points suggest that while AI is a powerful defensive tool, it is best used as an augmentation tool rather than a replacement tool.

TABLE V

RISK FRAMEWORK OF AI-DRIVEN THREAT DETECTION

Risk Category	Description	Potential Impact
Automation Bias	Excessive trust in system outputs	Reduced vigilance
Adversarial Manipulation	Crafted inputs mislead models	Detection failure
Model Drift	Accuracy declines over time	Misclassification
Alert Suppression	Filtering hides signals	Delayed response
Context Limitation	Algorithms lack situational awareness	Investigation gaps

III. PROPOSED SCHEME

A. Controlled Trust Scheme for AI-Driven Threat Detection

The increasing adoption of Artificial Intelligence in the field of cybersecurity has increased the ability of organizations to identify threats in a complex digital environment. AI-based detection systems have the ability to process a large amount of security data, identify anomalies in behavior, and raise an alert without any delay [13], [14]. However, due to the over-reliance on these automated systems, there is a possibility that organizations may accept the output of these systems without any critical evaluation of the output, which may lead to a situation of false sense of security.

To overcome this problem, the present study proposes a Controlled Trust Scheme (CTS) that seeks to strike a balance between machine intelligence and human intelligence. The proposed scheme suggests a multi-layered approach to analysis, wherein the security data obtained from the network, endpoint, and cloud environment is first consolidated to gain better understanding. The data is then analyzed by an AI-based detection system that uses behavioral analysis to identify potential threats. Instead of accepting the detection output as the final output, the system uses a confidence level that measures the probability of the threat.

The response actions are dependent on the confidence level of the alert. High-confidence alerts are immediately responded to by the automated response system, whereas alerts with moderate risk levels are evaluated by the analysts before taking any further course of action. Alerts with lower confidence levels are stored for behavioral analysis, which helps to identify potential long-term threats. The inclusion of structured validation in this process helps to overcome the risk of automation bias and emphasizes the need for critical evaluation of AI-based outputs [16].

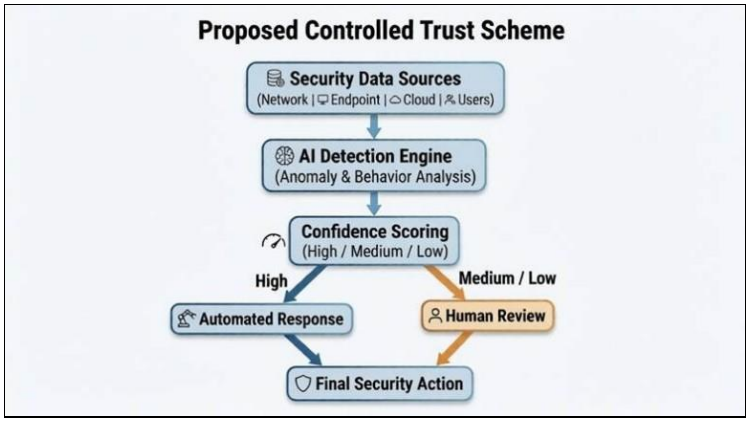

Fig.3. Controlled Trust Architecture

B. Security Impact and Operational Benefits of the Proposed Scheme

The integration of Artificial Intelligence has improved the capacity of organizations to detect threats in a complex digital environment. AI-powered detection tools have the potential to analyze a large amount of security data, identify anomalies in behavior, and generate alerts with minimal latency [13], [14]. However, over-reliance on the results of automated systems can result in a scenario where organizations are tempted to rely on the results of algorithms without sufficient verification, resulting in a false sense of security.To address this concern, this research work proposes a Controlled Trust Scheme (CTS), which balances the strengths of machines and human intervention. The strategy takes a multi-layered approach that begins with data aggregation from networks, endpoints, and cloud infrastructure to offer contextual awareness. The aggregated data is then analyzed by an AI-powered detection engine that examines behavioral patterns and identifies possible threats. Instead of offering the result in a conclusive manner, the system offers a confidence level that measures the probability of malicious activity. High-risk alerts marked with high confidence levels initiate automatic containment, whereas alerts marked with moderate risk levels require validation by analysts prior to taking any action. Alerts with low confidence levels are archived for behavioral correlation to enable long-term threat analysis. The integration of structured verification in the detection process prevents the possibility of automation bias and encourages critical thinking about AI-powered alerts [16].

C. Architectural Components of the Controlled Trust Scheme

Controlled Trust Scheme (CTS) is a system that aims at reducing the risk connected with the increased reliance on the use of Artificial Intelligence when performing security activities. Instead of relying on security decisions provided by artificial intelligence, the framework provides specific validation procedures which ensure balance between artificial and human intelligence. It contains five interrelated modules which contribute to enhancing trust calibration and minimizing the likelihood of having false security confidence.

1. Security Data Collection Layer

The first module of the framework is aimed at gathering security-related data from various organizational assets. In particular, such assets as network devices, endpoints, cloud, authentication, users and applications are utilized in order to get a complete picture of security state of an organization.

The performance of threat detection algorithms relies on the availability of high-quality data. That is why the data collection module contributes to gathering security events from different sources and ensuring better visibility.

2. AI Detection Engine

The AI Detection Engine conducts behavioral analysis, detects anomalies, and classifies threats. The difference between the AI Detection Engine and other traditional security solutions is that the engine detects deviation from expected behaviors and not only uses known signatures.

The engine analyzes the security telemetry received and looks for the signs that can be linked to malicious activities. These signs can include odd logins, suspicious communications in the network, privilege escalation attempts, and suspicious accesses to data.

The main task of this part of the system is to increase the efficiency of detection while allowing early detection of emerging threats which do not have signatures yet.

3. Confidence Scoring Module

One of the unique things about the Controlled Trust Scheme is the inclusion of confidence scoring. In contrast to other schemes where all alerts are treated equally, here a value of confidence in the prediction of AI is calculated.

The confidence value depends on several parameters, including:

- Detection history
- Severity of the activity detected
- Consistency of behavior
- Correlation with earlier incidents
- Reliability of the data source

4. Human Validation Layer

The Human Validation Layer is a safety measure to counter the effect of automation bias. Security alerts with medium levels of confidence are submitted to the analysts for evaluation prior to carrying out the recommended actions.

Human validation entails:

- Operational context
- Business impact
- Likely threat
- Attack trends
- Risk threshold

It serves as a means of ensuring that important security decisions are not solely dependent on algorithms and prevents overreliance on automation.

5. Feedback and Continuous Learning Layer

The last element of the framework enables the continuous improvement of the decision-making process. Analyst decisions are noted and incorporated into the future training of the model.

This feedback method helps to:

- Decrease false alarms
- Decrease missed alerts
- Improve confidence scores calibration
- Account for evolving threats
- Improve detection accuracy

With the feedback loop, the model continues improving together with the organization environment and cyber threats.

Incorporation of the five elements above creates a comprehensive cybersecurity model where the Artificial Intelligence works efficiently while the human input takes care of the accountability and transparency.

D. Operational Workflow of the Controlled Trust Scheme

The Controlled Trust Model uses a well-defined workflow process, which utilizes automated intelligence alongside human oversight. The workflow is aimed at making sure the level of trust in the results of AI is commensurate with the reliability of the evidence.

First, security telemetry information is gathered continuously from the infrastructure of the organization. Data from endpoints,

clouds, networks, authentications, and users' actions is gathered in one centralized monitoring environment.

After the gathering of data, processing happens through the AI Detection Engine. It undertakes behavioral analysis and anomaly detections on the data. The detected anomalies in this stage receive preliminary threat ratings.

Finally, the Confidence Scoring Module rates the reliability of each threat. The confidence score is calculated from such elements as past performance of the model, severity level of the threat, contextual consistency, and behavioral correlation.

High-confidence alerts are those which possess confidence scores higher than some preset thresholds. Given the presence of obvious signs of malicious activities, automated responses including account suspension, device isolation, and network containment may be performed right away.

Moderate-confidence alerts are sent to Human Validation Layer. In this layer, analysts analyze the cases, make judgments on the basis of contextual data, and perform further response actions where applicable. At this stage, the probability of automation bias is significantly reduced.

Low-confidence alerts are not dismissed; rather, they are stored and monitored in order to be correlated later on with newly discovered threat indicators. It allows detecting persistent campaigns even if they do not show obvious signs of being harmful at once.

Once the investigation and response steps are done, analyst’s decisions go into the Feedback and Continuous Learning Layer and are used in improving future models.

The work-flow ensures that security decisions will be made on the basis of both algorithmic analysis and contextual human judgment, which makes the probability of false sense of security lower.

E. Quantitative Evaluation Framework

While the Controlled Trust scheme is offered as a conceptual model, its efficiency may be determined through a set of cybersecurity performance metrics. They allow for an organized evaluation of how successful the model is in increasing detection quality without an overreliance on Artificial Intelligence.

TABLE VI

PERFORMANCE METRICS FOR CONTROLLED TRUST SCHEME EVALUATION

Metric	Description	Desired Outcome
Detection Accuracy	Percentage of correctly identified threats	High
False Positive Rate	Benign activities incorrectly flagged as threats	Low
False Negative Rate	Malicious activities that remain undetected	Low
Mean Detection Time	Average time required to identify threats	Low
Analyst Verification Rate	Percentage of alerts requiring analyst review	Balanced
Risk Reduction Index	Reduction in successful security incidents	High
Alert Reliability Score	Consistency of AI-generated alerts	High
Decision Confidence Accuracy	Accuracy of confidence score assignments	High

The adoption of the Controlled Trust Scheme would help enhance reliability by minimizing the effects of overreliance on the results generated by Automated Intelligence but maintaining the benefits of Artificial Intelligence.

A decrease in false positives would minimize the stress of the analysts and increase efficiency in operation. The same way, a decrease in false negatives would increase chances of detecting a threat before considerable damage can be done.
Confidence scoring would increase the quality of the decisions made as any suspicious alert will be further investigated. In addition, human validation will minimize the issue of automation bias and therefore make the system resilient to attacks.

Empirical validation of the scheme can be done in the future through the use of publicly available cybersecurity datasets such as UNSW-NB15, CICIDS2017, CSE-CIC-IDS2018 among others. The comparison between conventional AI system and Controlled Trust Scheme would give quantitative evidence on the performance of the confidence-based approach to security decision-making.

IV. ANALYSIS

Integration of Artificial Intelligence into cybersecurity operations enables companies to enhance their threat detection capabilities significantly. AI-based detectors are capable of processing large amounts of security data, spotting behavioral anomalies, and issuing alerts much faster than any human could. Despite those benefits, the growing use of intelligent systems comes with a number of organizational and technical challenges contributing to a false sense of security.

The first challenge associated with intelligent systems is automation bias. The more often security analysts see successful detections carried out with the help of algorithms, the more confidence they gain in algorithmic recommendations. As a result, validation of the warnings becomes less likely, which could lead to the neglecting of some vital indicators. Ultimately, the company may think that no alerts equal no threats.

Another crucial risk is related to the possibility of adversaries' manipulations. The modern machine learning systems base their decision-making on patterns identified in the training data. Adversaries can take advantage of that trait of intelligent systems by designing the input which looks similar to the normal behavior but contains a hidden malicious meaning.

Model drift is yet another problem when implementing AI-powered threat detection tools. Every security environment changes constantly because of the introduction of new technologies, procedures, and methods of communication. Therefore, models trained using the historical data will become outdated with time unless they undergo frequent validation and updating. The change of effectiveness occurs gradually and, hence, is hard to notice.

The problem of explainability is yet another one that leads to false security confidence. Contemporary machine learning models tend to be quite complex black boxes that do not provide transparency in the decision-making process to security specialists. If analysts do not know the reason for generating certain alerts, they would not be able to assess them effectively.

Controlled Trust Scheme provides the solution to all these problems with its structured trust management approach. The scoring of confidence means that the alerts are evaluated in terms of their trustworthiness rather than accepted without question. Human validation of alerts brings context to the process and helps to avoid automation bias.

TABLE VII

EXPECTED IMPACT OF CONTROLLED TRUST SCHEME IMPLEMENTATION

Security Dimension	AI-Only Environment	CTS Environment
Automation Bias	High	Reduced
Human Oversight	Limited	Structured
Alert Reliability	Moderate	Improved
Decision Transparency	Limited	Enhanced
False Security Confidence	High Risk	Controlled
Analyst Accountability	Reduced	Improved
Long-Term Adaptability	Moderate	Higher
Decision Reliability	Moderate	High

The analysis shows that effective cybersecurity is not only based on the level of detection but also on the way in which the trust management in automatic systems is performed. Artificial Intelligence is still an important defense instrument; nonetheless, its advantages cannot be realized to their full extent without proper governance, validation, and human control measures. The

Controlled Trust Scheme is a good example of such an approach.

V. LIMITATIONS

In this research, a conceptual method is used to analyze the risks involved in AI-powered threat detection. Although the model is validated by existing literature on cybersecurity, it would be more useful if it were empirically tested in practical settings. Moreover, the dynamic nature of threats could affect the relevance of risks as detection systems become more advanced. The maturity level of organizational security also differs from one industry to another, and hence the level of dependence on AI-powered systems may not be the same. Moreover, this research is mainly concerned with the dynamics of the detection phase, while the overall environment of cybersecurity resilience is influenced by other factors such as response planning and user awareness.

VI. CONCLUSION

The application of AI threat detection has greatly improved the efficiency of cybersecurity by enabling rapid threat detection and the scalability of large systems. However, the overdependence on the advancements of technology may not always guarantee foolproof security. Overemphasizing the advancements of technology may lead to a situation where a false sense of security is created, which may lead to the possibility of sophisticated threats being undetected. By assessing the current state of literature and applying comparative knowledge, the research study has been able to conclude that automation bias, adversarial attacks, and model limitations are essential components in the creation of a false sense of security. The new Controlled Trust Scheme will remove this issue by ensuring confidence scoring and human assessment, which will enable more informed security decision-making.

Finally, there must be a symbiotic relationship between human intelligence and artificial intelligence for the success of effective cybersecurity. By recognizing AI as a complement to their analytical work, rather than a panacea, organizations can better prepare for the ever-changing landscape of the cyber threat environment.

REFERENCES

[1] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. Upper Saddle River, NJ, USA: Prentice Hall, 2010.

[2] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[3] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in *Proc. IEEE Symposium on Security and Privacy*, 2010.

[4] Verizon, "2023 Data Breach Investigations Report (DBIR)," Verizon Enterprise, 2023.

[5] E. B. Karbab, M. Debbabi, A. Derhab and D. Mouheb, "MalDozer: Automatic framework for Android malware detection using deep learning," *Digital Investigation*, vol. 24, pp. S48–S59, 2018.

[6] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive dataset for network intrusion detection systems," in *Proc. Military Communications and Information Systems Conference (MilCIS)*, 2015.

[7] M. A. Ferrag, L. Maglaras, H. Janicke, J. Jiang and T. Shu, "Authentication protocols for Internet of Things: A comprehensive survey," *Security and Communication Networks*, 2017.

[8] A. Javaid, Q. Niyaz, W. Sun and M. Alam, "A deep learning approach for network intrusion detection system," in *Proc. EAI International Conference on Bio-inspired Information and Communications Technologies (EAI ICST)*, 2016.

[9] S. Axelsson, "Intrusion detection systems: A survey and taxonomy," Technical Report, Chalmers University of Technology, 2000.

[10] G. Creech and J. Hu, "A semantic approach to host-based intrusion detection systems using contiguous and discontiguous system call patterns," *IEEE Transactions on Computers*, 2014.

[11] D. E. Denning, "An intrusion-detection model," *IEEE Transactions on Software Engineering*, vol. SE-13, no. 2, pp. 222–232, 1987.

[12] K. Scarfone and P. Mell, "Guide to intrusion detection and prevention systems (IDPS)," *NIST Special Publication 800-94*, 2007.

[13] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, vol. 84, pp. 317–331, 2018.

[14] N. Papernot*et al.*, "The limitations of deep learning in adversarial settings," in *Proc. IEEE European Symposium on Security and Privacy*, 2016.

[15] A. Kurakin, I. Goodfellow and S. Bengio, "Adversarial examples in the physical world," in *Proc. International Conference on Learning Representations (ICLR) Workshop*, 2017.

[16] M. Skitka, K. Mosier and M. Burdick, "Does automation bias decision-making?" *International Journal of Human-Computer Studies*, vol. 51, no. 5, pp. 991–1006, 1999.

[17] B. Lyell and D. Coiera, "Automation bias and verification complexity: A systematic review," *Journal of the American Medical Informatics Association*, 2017.

[18] ENISA, "ENISA Threat Landscape 2023," European Union Agency for Cybersecurity, 2023.

[19] U.S. Government Accountability Office, "Data protection: Actions taken by Equifax and federal agencies," GAO Report, 2018.

[20] P. Cichonski, T. Millar, T. Grance and K. Scarfone, "Computer Security Incident Handling Guide," *NIST Special Publication 800-61 Rev. 2*, 2012.