# Detecting Malicious Accounts in Cyberspace: Enhancing Security in ChatGPT and Beyond

**Said A. Salloum**

## 1 Introduction

In the rapidly evolving digital age, cybersecurity has become an indispensable aspect of safeguarding both personal data and the operational integrity of online platforms. The pervasive use of the internet in various facets of life underscores the critical importance of robust cybersecurity measures. Studies indicate that as our reliance on digital technologies grows, so too does the sophistication and frequency of cyber-attacks [1].

Central to the challenges faced in the realm of cybersecurity is the proliferation of malicious accounts in cyberspace. These accounts, which can range from automated bots to sophisticatedly disguised fake profiles, represent a multifaceted threat. They are known to engage in activities that undermine the authenticity of online interactions, spread misinformation, and execute fraudulent schemes [2].

The impact of these malicious entities extends across various digital platforms, notably affecting advanced AI systems like ChatGPT. In the context of AI-driven platforms, these accounts pose unique challenges. They can manipulate conversational dynamics, skew AI learning processes, and compromise the reliability of AI-generated content. This not only deteriorates the user experience but also raises significant concerns about the trustworthiness and security of these platforms [3].

Figure 1 provides a vivid conceptual representation of the pervasive and disruptive impact of malicious entities across a wide range of digital platforms. The illustration captures a network of interconnected platforms, including social media, online forums, and AI chat interfaces, highlighting the ubiquity of digital communication in modern society. These platforms are shown being infiltrated by symbols representing

S. A. Salloum (✉)
Health Economic and Financing Group, University of Sharjah, Sharjah, UAE
e-mail: ssalloum@sharjah.ac.ae

School of Science, Engineering, and Environment, University of Salford, Salford, UK



A. Al-Marzouqi et al. (eds.), *Artificial Intelligence in Education: The Power and Dangers of ChatGPT in the Classroom*, Studies in Big Data 144,
https://doi.org/10.1007/978-3-031-52280-2_42

various forms of malicious entities—such as bots, trolls, and fake profiles—which are depicted as causing disturbances and disruptions within the network. The symbols for malicious accounts are strategically placed to demonstrate their ability to blend in and yet distinctly disrupt the normal flow of digital communication and interaction. This visualization underscores the stealth and sophistication with which these entities operate, aligning with findings from [4] which detail the evolving nature of digital threats. The disruption caused by these entities is not just isolated to one platform but is shown to have a cascading effect, illustrating how vulnerabilities in one area can lead to broader compromises in the digital ecosystem. In the context of AI chat interfaces, such as ChatGPT, the image reflects the specific challenges these platforms face, including the manipulation of conversational dynamics and the potential skewing of AI learning processes. This aspect of the illustration aligns with studies like [5], which discuss the unique vulnerabilities and security considerations for AI-driven platforms.

This paper aims to delve into the intricate landscape of detecting and mitigating the presence of malicious accounts in cyberspace. We will explore the array of current detection methodologies, ranging from traditional cyber security tactics to innovative AI-driven approaches. The paper will critically evaluate the effectiveness of these strategies in enhancing the security measures of digital platforms, with a particular focus on AI-driven systems like ChatGPT. Additionally, we will examine the broader implications of these cybersecurity measures, considering their potential impact on user privacy, data integrity, and the ethical dimensions of AI moderation. The scope of this paper encompasses a comprehensive review of existing literature, an analysis of case studies pertaining to current cybersecurity practices, and a forward-looking discussion on emerging trends and challenges in this dynamic field.

**Fig. 1** Conceptual representation of the impact of malicious entities on digital platforms

## 2 Understanding Malicious Accounts

### *2.1 Definition and Types of Malicious Accounts*

In the realm of cybersecurity, the term 'malicious accounts' encompasses a range of deceptive and harmful online entities, each with distinct characteristics and objectives. Bots, a prevalent form of malicious accounts, are automated programs designed to mimic human actions online. They are often employed for spamming, spreading misinformation, or inflating social media metrics, as outlined in [6–8]. Trolls represent another category, typically human-operated accounts that aim to provoke and disrupt online conversations through inflammatory or deceptive content. Studies such as [9–11] delve into their impact on social media discourse and user experience.

Fake accounts, which include both fully fabricated profiles and those impersonating real individuals, are particularly insidious. These accounts are used for a range of malicious purposes, from spreading false information to phishing scams. Research by [12] provides insights into their detection and the challenges they pose to digital platform integrity. Each of these account types represents a unique threat to the online ecosystem, necessitating tailored detection and mitigation strategies. Understanding their distinct characteristics is crucial for effective cybersecurity measures, as they exploit different aspects of online platforms and user interactions.

### *2.2 Motivations and Methods Used by These Accounts*

Malicious accounts in cyberspace are driven by a diverse range of motivations, employing various methods to achieve their objectives [13]. At the core, many of these entities are designed to manipulate, deceive, or disrupt online interactions. For instance, bots, programmed for automated tasks, are often used for spreading misinformation or amplifying social media content to manipulate public opinion or distort online discussions [14]. Their methods range from mass-posting similar messages to more sophisticated interactions that mimic human behavior. Trolls, typically human operators behind anonymous accounts, primarily aim to provoke or distress others for personal amusement or to push a specific agenda. The methods employed by trolls include posting inflammatory comments, engaging in harassment, and creating divisive content, as highlighted in studies like [15]. Fake accounts, on the other hand, may have more deceptive goals, such as phishing, identity theft, or spreading false information. These accounts often imitate real users or organizations, leveraging their perceived authenticity to mislead and exploit other users [16]. Understanding the motivations and methods of these malicious entities is crucial for developing effective countermeasures. Each type of malicious account presents unique challenges, requiring tailored approaches for detection and mitigation. This understanding is not only important for cybersecurity professionals but also for regular users to recognize and safeguard against such threats in their online interactions.

### 2.3 The Impact of Malicious Accounts on User Experience and Platform Integrity

The presence of malicious accounts in cyberspace significantly impacts both user experience and the integrity of digital platforms. These accounts, through various means, erode the trust and usability of online ecosystems. Bots, for instance, can flood platforms with spam or misinformation, leading to a polluted information environment. This can diminish user experience by cluttering feeds with irrelevant or deceptive content, as explored in [17]. Their activity can also skew analytics, giving false impressions of popularity or consensus, which is particularly detrimental for platforms relying on user engagement metrics for decision-making or advertising purposes. Trolls negatively impact user experience by creating a hostile or toxic online environment. Their behavior can lead to harassment and cyberbullying, causing distress to individual users and often leading to reduced participation or complete withdrawal from the platform [18]. This not only affects individual well-being but also the overall quality of discourse on the platform.

Fake accounts pose a direct threat to platform integrity and user security. By impersonating legitimate users or entities, they can engage in phishing attacks, scamming users, or spreading false information, thereby compromising the authenticity and reliability of the platform [19, 20]. The presence of these accounts can lead to a lack of trust in the platform, as users become uncertain about the genuineness of the interactions they have and the content they consume.

The collective impact of these accounts is a significant challenge for platform operators, as it undermines user trust, engagement, and satisfaction—all of which are crucial for the long-term viability and success of digital platforms. Addressing these issues requires not only technical solutions but also a consideration of the broader social and ethical implications of these malicious activities.

## 3 Current Detection Approaches

### 3.1 Overview of Traditional Detection Methods

Traditional methods for detecting malicious accounts in cyberspace have primarily focused on identifying patterns and anomalies that differentiate these accounts from legitimate users. One common approach is IP tracking, which involves monitoring the IP addresses from which accounts are accessed. This method can help identify accounts that are being operated from known sources of malicious activity or through proxies typically used to mask illicit activities [21]. Another widely used technique is the analysis of account activity. This includes examining login patterns, posting frequencies, and interaction styles. For instance, accounts that exhibit non-human behavior, such as posting at superhuman speeds or exhibiting repetitive patterns, can be flagged as potential bots or fake accounts [22, 23]. These traditional methods,

while effective in certain contexts, have limitations. Malicious actors have become adept at evading detection by mimicking human behaviors or using sophisticated methods to mask their true IP addresses. Consequently, while these methods form an essential part of the cybersecurity toolkit, they are increasingly being supplemented with more advanced techniques, particularly those leveraging artificial intelligence and machine learning algorithms.

### *3.2 Machine Learning and AI in Detecting Malicious Accounts*

The integration of Machine Learning (ML) and Artificial Intelligence (AI) has revolutionized the detection of malicious accounts in cyberspace, offering more dynamic and sophisticated methods compared to traditional techniques. ML algorithms are particularly adept at pattern recognition, allowing them to identify subtle behaviors and characteristics indicative of malicious accounts that might be overlooked by human analysts or simpler automated systems. For instance, ML models can be trained on large datasets to recognize patterns typical of bot activities, such as the timing and frequency of posts, or the nature of interactions with other users [24–27]. AI, especially when it involves Deep Learning techniques, can analyze not just the metadata of accounts but also the content of the interactions. This allows for the identification of more complex behaviors, such as the dissemination of misinformation or coordinated inauthentic behaviors that are characteristic of sophisticated fake accounts or state-sponsored trolling operations [28–33]. These technologies are not without their challenges. The evolving nature of malicious tactics means that detection systems must continuously learn and adapt, a process that can be resource-intensive. Additionally, the risk of false positives, where legitimate accounts are mistakenly identified as malicious, poses significant ethical and operational considerations [34]. Despite these challenges, the use of ML and AI in this domain represents a significant advancement in cybersecurity efforts. They offer scalability and efficiency in monitoring and analyzing vast amounts of data, an essential capability given the sheer scale of modern digital platforms.

### *3.3 The Role of User Reporting and Community Management*

In the ecosystem of digital platform security, user reporting and community management play pivotal roles in identifying and mitigating the risks posed by malicious accounts. User reporting is a crucial first line of defense, enabling the platform's community to flag suspicious or harmful content or behavior. This grassroots level of surveillance harnesses the collective vigilance of users, often catching anomalies that

automated systems might miss. Studies have shown that user reports can be instrumental in identifying coordinated disinformation campaigns or isolated instances of harassment, underscoring the importance of community engagement in maintaining platform integrity [2, 32]. Community management, on the other hand, involves a more proactive approach by platform administrators or designated moderators. This includes setting and enforcing community guidelines, monitoring discussions, and responding to user reports. Effective community management not only deters malicious activities through policy enforcement but also fosters a sense of safety and trust among users. It has been observed that platforms with active and visible community management have lower incidences of malicious activities, as they create an environment less conducive to the objectives of such accounts [35]. The synergy of user reporting and community management forms a comprehensive approach to platform security. It balances automated detection systems with human insight and judgment, essential in a landscape where malicious actors continually evolve their tactics. However, this approach also raises challenges, particularly in ensuring timely and appropriate responses to user reports and in managing the potential biases of human moderators [2].

## 4 Application to ChatGPT and Other AI Platforms

### *4.1 Specific Vulnerabilities of AI Platforms like ChatGPT to Malicious Accounts*

AI platforms like ChatGPT are uniquely vulnerable to certain types of malicious accounts, mainly due to their reliance on user interactions and data for learning and response generation. One of the primary vulnerabilities is the potential for data poisoning, where malicious accounts feed misleading or harmful information to the AI, influencing its learning process and output. This type of attack can subtly skew the AI's language models, leading to biased or inappropriate responses [36]. Another vulnerability is the exploitation of AI's response mechanisms by malicious accounts to amplify misinformation or harmful content. Since AI platforms like ChatGPT are designed to engage in natural and relevant conversations, they can be manipulated into responding or interacting with content that furthers the agenda of these accounts [37].

Additionally, AI platforms may face challenges in distinguishing between legitimate user interactions and those orchestrated by malicious accounts, especially when these accounts employ sophisticated tactics to mimic genuine behavior. This difficulty can lead to inadequate responses to harmful content or the unintentional dissemination of such content, compromising the integrity and trustworthiness of the AI platform [38]. These vulnerabilities necessitate a multifaceted approach to AI

platform security, combining advanced algorithmic defenses with oversight mechanisms to ensure the AI's interactions and learning processes remain aligned with ethical and factual standards.

## *4.2 How Current Detection Approaches Can Be Adapted for ChatGPT?*

Adapting current detection approaches for ChatGPT involves several strategies that align with the unique characteristics and operational contexts of AI-driven conversation platforms. Machine learning algorithms, which are already a core component of ChatGPT, can be fine-tuned to identify patterns indicative of malicious account activities. This includes training models on datasets that capture the nuances of such activities, ranging from spam and troll behavior to more sophisticated misinformation campaigns [39]. Natural language processing (NLP) techniques, integral to ChatGPT's functionality, can be leveraged to analyze conversational contexts and identify potential malicious interactions. Advanced NLP can discern subtleties in language that are indicative of harmful content or deceptive practices, a method that has shown effectiveness in other digital platforms [40–42].

In addition to these AI-centric methods, user feedback mechanisms within ChatGPT can play a crucial role. Implementing robust user reporting features and feedback loops allows for the gathering of valuable user insights, which can be used to continually refine detection algorithms and response strategies [43].

Furthermore, incorporating a multi-layered security approach that combines these AI-driven methods with traditional cybersecurity practices, such as IP tracking and account verification, can enhance the overall resilience of ChatGPT against malicious accounts [44]. The adaptation of these detection methods for ChatGPT underscores the need for a dynamic and evolving approach to platform security, one that is capable of responding to the continuously changing tactics of malicious actors in the digital space.

## *4.3 Potential for AI-Driven Detection Methods to Improve Platform Security*

The potential of AI-driven detection methods in enhancing the security of digital platforms is significant, particularly as cyber threats become more sophisticated. These AI methods, primarily rooted in advanced machine learning and deep learning techniques, offer the ability to analyze vast amounts of data quickly and efficiently, a capability that is essential in identifying and mitigating cyber threats in real-time. Machine learning algorithms, for instance, can be trained to recognize patterns and

anomalies that signify malicious activities, such as unusual login behaviors or atypical content dissemination patterns, offering a level of analysis that is unfeasible for human monitors alone [45–47].

Deep learning, a subset of machine learning, provides even greater potential due to its ability to process and analyze complex datasets, including unstructured data like text and images. This capability is particularly valuable in detecting sophisticated cyber threats that traditional methods might miss, such as subtle phishing attempts or advanced social engineering tactics [48–50].

Furthermore, the integration of natural language processing (NLP) in AI-driven security systems enables the detection of nuanced malicious activities in textual content, such as hate speech, misinformation, or harmful propaganda. This approach is increasingly relevant in the context of social media and communication platforms, where such content can have a widespread impact [51]. The adaptability of AI-driven methods is another key advantage. AI models can be continuously updated and retrained to keep up with evolving cyber threats, ensuring that platform security measures remain effective over time. However, the implementation of these AI technologies also requires careful consideration of privacy and ethical standards to avoid potential misuse or bias in the detection processes [52].

## 5 Future Trends and Challenges

### *5.1 Emerging Technologies and Methods in Malicious Account Detection*

The landscape of cybersecurity is continually evolving, with emerging technologies and methods playing a critical role in detecting malicious accounts. Among the most promising advancements is the application of Artificial Intelligence (AI) and Machine Learning (ML), which offer sophisticated analytical capabilities. AI and ML algorithms can process and learn from vast amounts of data, identifying patterns and anomalies indicative of malicious behavior that may elude traditional detection methods [53, 54].

Another emerging technology is the use of blockchain for security purposes. Blockchain's inherent properties, such as decentralization, transparency, and immutability, make it a potentially powerful tool in the fight against cyber threats, including the detection and prevention of fraudulent activities by malicious accounts [55]. Network analysis is also gaining traction as a method for detecting malicious accounts. By examining the connections and patterns of interactions between accounts, it becomes possible to identify coordinated malicious activities, such as botnets or troll farms, which often operate in networks [56, 57].

Furthermore, the development of advanced natural language processing (NLP) techniques enhances the ability to scrutinize the content generated by accounts, spotting signs of manipulation or harmful intent in textual data [58].

Additionally, the integration of biometric verification methods, such as facial recognition or voice analysis, is being explored as a means to authenticate users and flag accounts that exhibit signs of falsification [59]. These emerging technologies and methods collectively represent a paradigm shift in cybersecurity strategies, offering more proactive and comprehensive approaches to safeguarding digital platforms from the ever-evolving threat posed by malicious accounts.

## *5.2 The Evolving Nature of Malicious Activities and the Need for Adaptive Security Measures?*

The landscape of cyber threats is in a constant state of flux, with malicious activities continuously evolving in complexity and sophistication. This dynamic nature poses a significant challenge for cybersecurity, necessitating adaptive and forward-thinking security measures. The shift from relatively straightforward phishing attacks to more complex, multi-vector threats exemplifies this evolution. Cybercriminals now employ a range of tactics, including advanced social engineering, AI-driven attacks, and sophisticated malware, to exploit vulnerabilities in digital systems [60].

In response to these changing threats, the development and implementation of adaptive security measures have become paramount. Machine learning and artificial intelligence are at the forefront of this adaptive approach. These technologies enable continual learning from new data, allowing security systems to evolve alongside emerging threats [61]. Furthermore, the integration of threat intelligence platforms helps in gathering and analyzing information about new and existing threats, providing insights that inform security strategies and responses [62]. Another key component of adaptive security measures is the emphasis on proactive rather than reactive strategies. This involves anticipating potential security incidents and preparing defenses in advance, rather than merely responding to breaches after they occur. Techniques such as predictive analytics and risk assessment models are instrumental in this proactive approach [63]. However, the evolving nature of cyber threats also calls for a broader perspective that includes legal, ethical, and policy considerations. The development of comprehensive cybersecurity policies and adherence to ethical standards is crucial to ensure that adaptive security measures are not only effective but also respect user privacy and data rights [64].

The continuous evolution of malicious activities in the digital domain underscores the need for security measures that are equally dynamic and responsive, blending technological innovation with strategic foresight and ethical responsibility.

### 5.3 *Ethical Considerations and Privacy Concerns in Detection Strategies*

In the realm of cybersecurity, particularly in the detection of malicious accounts, ethical considerations and privacy concerns are paramount. The deployment of sophisticated detection strategies, while essential for security, often raises questions regarding user privacy and data protection. Advanced monitoring and data analysis techniques, such as deep packet inspection or extensive data mining, can inadvertently infringe on individual privacy rights. The balance between security and privacy is a subject of ongoing debate, with scholars like [65] emphasizing the need for a nuanced approach. Another ethical concern relates to the potential for biases in AI-driven detection systems. Machine learning models, if trained on biased or unrepresentative data sets, can lead to discriminatory outcomes, unfairly targeting certain groups or individuals. Studies such as [66] have highlighted the necessity for unbiased data and transparent algorithms to mitigate these risks.

Furthermore, the ethical use of user data in cybersecurity measures is an area of significant concern. The collection and analysis of user data for security purposes must comply with data protection regulations and ethical standards, a topic explored in-depth by [67–69]. This compliance is not only a legal obligation but also critical for maintaining user trust in digital platforms.

In addition, the use of intrusive detection methods can raise ethical questions about the extent of surveillance and monitoring that is acceptable. The work of [67] offers insights into ethical limitations and guidelines for such practices. Addressing these ethical considerations and privacy concerns is crucial for the development of effective and responsible cybersecurity strategies. It involves a careful balance between the need for security and the protection of individual rights, calling for ongoing dialogue and collaboration between technologists, ethicists, and policymakers.

## 6 Conclusion

In conclusion, this exploration into the detection of malicious accounts in cyberspace underscores a landscape marked by complexity and ever-evolving challenges. Key findings indicate that while traditional methods like IP tracking and account activity analysis remain foundational, the emergence of AI and ML technologies has significantly enhanced the capability to identify and mitigate cyber threats. The adaptation of these advanced methods to specific platforms, particularly AI-driven systems like ChatGPT, is crucial in addressing the unique vulnerabilities they face. However, as our understanding and technological capabilities expand, so too does the sophistication of malicious entities, necessitating a continuous innovation in detection methods to maintain robust platform security. The importance of this ongoing innovation cannot be overstated, as it represents not only a response to emerging threats but also a proactive approach to safeguarding digital ecosystems. The integration of

emerging technologies such as blockchain, network analysis, and biometric verification further illustrates the dynamic nature of cybersecurity strategies. Yet, as we advance in our technical prowess, the need to maintain a delicate balance becomes increasingly paramount. Balancing security with user privacy and platform usability is a complex but essential task, requiring a nuanced approach that respects individual rights while ensuring a safe and reliable online experience. This balance is not static but a dynamic equilibrium that must be constantly reassessed in the light of new technologies, user expectations, and the evolving nature of cyber threats. The future of digital platform security thus lies not only in technological advancement but also in our ability to ethically and responsibly integrate these innovations into the fabric of our digital lives.

## References


1. A. Hussain, A. Mohamed, S. Razali, A review on cybersecurity: challenges & emerging threats, in *Proceedings of the 3rd International Conference on Networking, Information Systems & Security*, pp. 1–7 (2020)
2. K.S. Adewole, N.B. Anuar, A. Kamsin, K.D. Varathan, S.A. Razak, Malicious accounts: dark of the social networks. J. Netw. Comput. Appl. **79**, 41–67 (2017)
3. B. Rathore, Future of AI & generation alpha: ChatGPT beyond boundaries. Eduzone Int. Peer Rev. Multidiscip. J. **12**(1), 63–68 (2023)
4. M.F. Ansari, B. Dash, P. Sharma, N. Yathiraju, The impact and limitations of artificial intelligence in cybersecurity: a literature review. Int. J. Adv. Res. Comput. Commun. Eng. (2022)
5. S.A. Yablonsky, AI-driven digital platform innovation. Technol. Innov. Manag. Rev. **10**(10) (2020)
6. S. Salloum, T. Gaber, S. Vadera, K. Sharan, A systematic literature review on phishing email detection using natural language processing techniques. IEEE Access (2022)
7. S. Salloum, T. Gaber, S. Vadera, K. Shaalan, A New English/Arabic parallel corpus for phishing emails. ACM Trans. Asian Low-Resource Lang. Inf. Process. (2023)
8. S. Salloum, T. Gaber, S. Vadera, K. Shaalan, Phishing website detection from URLs using classical machine learning ANN model, in *International Conference on Security and Privacy in Communication Systems*, pp. 509–523 (2021)
9. S. Bradshaw, L.-M. Neudert, P.N. Howard, Government responses to malicious use of social media. NATO Strat. Cent. Excell. Riga, Work. Pap. (2018)
10. E. Taylor, S. Walsh, S. Bradshaw, Industry responses to the malicious use of social media. Nato Strat. (2018)
11. S. Das Bhattacharjee, W.J. Tolone, V.S. Paranjape, Identifying malicious social media contents using multi-view context-aware active learning. Futur. Gener. Comput. Syst. **100**, 365–379 (2019)
12. S. Salloum, T. Gaber, S. Vadera, K. Shaalan, Phishing email detection using natural language processing techniques: a literature survey. Procedia Comput. Sci. **189**, 19–28 (2021)
13. I. Akour, N. Alnazzawi, R. Alfaisal, S.A. Salloum, Using classical machine learning for phishing websites detection from URLS
14. P.B. Brandtzaeg, A. Følstad, Why people use chatbots, in *Internet Science: 4th International Conference, INSCI 2017, Thessaloniki, Greece, November 22–24, 2017, Proceedings 4*, pp. 377–392 (2017)
15. M. Tomaiuolo, G. Lombardo, M. Mordonini, S. Cagnoni, A. Poggi, A survey on troll detection. Futur. Inter. **12**(2), 31 (2020)

16. A. Pathak, *An analysis of various tools, methods and systems to generate fake accounts for social media* (Northeast. Univ, Boston, Massachusetts December, 2014)
17. Z. Gilani, R. Farahbakhsh, J. Crowcroft, Do bots impact Twitter activity?, in *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 781–782 (2017)
18. P. Koncar, S. Walk, D. Helic, M. Strohmaier, Exploring the impact of trolls on activity dynamics in real-world collaboration networks, in *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 1573–1578 (2017)
19. S. Khaled, N. El-Tazi, H.M.O. Mokhtar, Detecting fake accounts on social media. IEEE Intern. Conf. Big Data (Big Data) **2018**, 3672–3681 (2018)
20. L. Caruccio, D. Desiato, G. Polese, Fake account identification in social networks. IEEE Intern. Conf. Big Data (Big Data) **2018**, 5078–5085 (2018)
21. Y. Jin, Z.-L. Zhang, K. Xu, F. Cao, S. Sahu, Identifying and tracking suspicious activities through IP gray space analysis, in *Proceedings of the 3rd annual ACM workshop on Mining network data*, pp. 7–12 (2007)
22. S.S. Tirumala, H. Sathu, V. Naidu, Analysis and prevention of account hijacking based incidents in cloud environment, in *2015 international Conference on Information Technology (ICIT)*, pp. 124–129 (2015)
23. P. Béguin, Taking activity into account during the design process. Activités **4**(4–2) (2007)
24. I. Dimitriadis, K. Georgiou, A. Vakali, Social botomics: A systematic ensemble ml approach for explainable and multi-class bot detection. Appl. Sci. **11**(21), 9857 (2021)
25. A. Abou Daya, M.A. Salahuddin, N. Limam, R. Boutaba, BotChase: Graph-based bot detection using machine learning. IEEE Trans. Netw. Serv. Manag. **17**(1), 15–29 (2020)
26. A. Abou Daya, M.A. Salahuddin, N. Limam, R. Boutaba, A graph-based machine learning approach for bot detection, in *2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, pp. 144–152 (2019)
27. S. Miller, C. Busby-Earle, The role of machine learning in botnet detection, in *2016 11th international conference for internet technology and secured transactions (icitst)*, pp. 359–364 (2016)
28. S. Kudugunta, E. Ferrara, Deep neural networks for bot detection. Inf. Sci. (Ny) **467**, 312–322 (2018)
29. K. Hayawi, S. Saha, M.M. Masud, S.S. Mathew, M. Kaosar, Social media bot detection with deep learning methods: a systematic review. Neural Comput. Appl. **35**(12), 8903–8918 (2023)
30. E. Arin, M. Kutlu, Deep learning based social bot detection on twitter. IEEE Trans. Inf. Forensics Secur. **18**, 1763–1772 (2023)
31. M. Rabbani et al., A review on machine learning approaches for network malicious behavior detection in emerging technologies. Entropy **23**(5), 529 (2021)
32. M. Brundage et al., The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv Prepr. arXiv1802.07228* (2018)
33. J.N. Paredes, G.I. Simari, M.V. Martinez, M.A. Falappa, Detecting malicious behavior in social platforms via hybrid knowledge-and data-driven systems. Futur. Gener. Comput. Syst. **125**, 232–246 (2021)
34. O. Ajibuwa, B. Hamdaoui, A.A. Yavuz, A survey on AI/ML-driven intrusion and misbehavior detection in networked autonomous systems: techniques, challenges and opportunities. *arXiv Prepr. arXiv2305.05040* (2023)
35. Q. Gong et al., Detecting malicious accounts in online developer communities using deep learning, in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 1251–1260 (2019)
36. J. Shen, M. Xia, AI data poisoning attack: manipulating game AI of Go. *arXiv Prepr. arXiv2007.11820* (2020)
37. G. Petropoulos, The dark side of artificial intelligence: manipulation of human behaviour. *Bruegel-Blogs*, p. NA-NA (2022)
38. M. Bhattacharya, S. Roy, S. Chattopadhyay, A.K. Das, S. Shetty, A comprehensive survey on online social networks security and privacy issues: Threats, machine learning-based solutions, and open challenges. Secur. Priv. **6**(1), e275 (2023)

39. A.S. George, A.S.H. George, A review of ChatGPT AI's impact on several business sectors. Partners Univers. Int. Innov. J. **1**(1), 9–23 (2023)
40. S. Gharge, M. Chavan, An integrated approach for malicious tweets detection using NLP," in *2017 international conference on inventive communication and computational technologies (ICICCT)*, pp. 435–438 (2017)
41. H. Yang, Q. He, Z. Liu, Q. Zhang, Malicious encryption traffic detection based on NLP. Secur. Commun. Networks **2021**, 1–10 (2021)
42. M. Mimura, H. Miura, Detecting unseen malicious VBA macros with NLP techniques. J. Inf. Process. **27**, 555–563 (2019)
43. B. Porter, F. Grippa, A platform for AI-enabled real-time feedback to promote digital collaboration. Sustainability **12**(24), 10243 (2020)
44. J. Alves-Foss, C. Taylor, P. Oman, A multi-layered approach to security in high assurance systems, in *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the*, pp. 10 (2017)
45. S.A. Salloum, M. Alshurideh, A. Elnagar, K. Shaalan, Machine learning and deep learning techniques for cybersecurity: a review, in *Joint European-US Workshop on Applications of Invariance in Computer Vision*, pp. 50–57 (2020)
46. D. Dasgupta, Z. Akhtar, S. Sen, Machine learning in cybersecurity: a comprehensive survey. J. Def. Model. Simul. **19**(1), 57–106 (2022)
47. Y. Xin et al., Machine learning and deep learning methods for cybersecurity. IEEE Access **6**, 35365–35381 (2018)
48. S. Mahdavifar, A.A. Ghorbani, Application of deep learning to cybersecurity: a survey. Neurocomputing **347**, 149–176 (2019)
49. P. Dixit, S. Silakari, Deep learning algorithms for cybersecurity applications: a technological and status review. Comput. Sci. Rev. **39**, 100317 (2021)
50. Y.N. Imamverdiyev, F.J. Abdullayeva, Deep learning in cybersecurity: challenges and approaches. Int. J. Cyber Warf. Terror. **10**(2), 82–105 (2020)
51. A. Doan, N. England, T. Vitello, Online review content moderation using natural language processing and machine learning methods: 2021 systems and information engineering design symposium (SIEDS). Syst. Inform. Eng. Des. Symp. (SIEDS) **2021**, 1–6 (2021)
52. N.M. Safdar, J.D. Banja, C.C. Meltzer, Ethical considerations in artificial intelligence. Eur. J. Radiol. **122**, 108768 (2020)
53. F. Kamoun, F. Iqbal, M.A. Esseghir, T. Baker, AI and machine learning: a mixed blessing for cybersecurity, in *2020 International Symposium on Networks, Computers and Communications (ISNCC)*, pp. 1–7 (2020)
54. K. Bresniker, A. Gavrilovska, J. Holt, D. Milojicic, T. Tran, Grand challenge: applying artificial intelligence and machine learning to cybersecurity. Computer (Long. Beach. Calif) **52**(12), 45–52 (2019)
55. P. Zhuang, T. Zamir, H. Liang, Blockchain for cybersecurity in smart grid: A comprehensive survey. IEEE Trans. Ind. Inform. **17**(1), 3–19 (2020)
56. C. Modi, D. Patel, B. Borisaniya, H. Patel, A. Patel, M. Rajarajan, A survey of intrusion detection techniques in Cloud. J. Netw. Comput. Appl. **36**(1), 42–57 (Jan.2013)
57. B. Mukherjee, L.T. Heberlein, K.N. Levitt, Network intrusion detection. IEEE Netw. **8**(3), 26–41 (1994)
58. A.A. Sattikar, R.V. Kulkarni, Natural language processing for content analysis in social networking. Int. J. Eng. Invent. **1**(4), 6–9 (2012)
59. M. Adán, A. Adán, A.S. Vázquez, R. Torres, Biometric verification/identification based on hands natural layout. Image Vis. Comput. **26**(4), 451–465 (2008)
60. S. Ganapati, M. Ahn, C. Reddick, Evolution of cybersecurity concerns: a systematic literature review, in *Proceedings of the 24th Annual International Conference on Digital Government Research*, pp. 90–97 (2023)
61. E. Iturbe, E. Rios, A. Rego, N. Toledo, Artificial Intelligence for next generation cybersecurity: The AI4CYBER framework, in *Proceedings of the 18th International Conference on Availability, Reliability and Security*, pp. 1–8 (2023)

62. D.P.F. Möller, Threats and threat intelligence, in *Guide to Cybersecurity in Digital Transformation: Trends, Methods, Technologies, Applications and Best Practices*, Springer, pp. 71–129 (2023)
63. N. Sun et al., Cyber Threat intelligence mining for proactive cybersecurity defense: a survey and new perspectives, *IEEE Commun. Surv. Tutorials* (2023)
64. N. Allahrakha, Balancing cyber-security and privacy: legal and ethical considerations in the digital age. Leg. Issues Digit. Age **4**(2), 78–121 (2023)
65. B.F.G. Fabrègue, A. Bogoni, Privacy and security concerns in the smart city. Smart Cities **6**(1), 586–613 (2023)
66. K. Michael, R. Abbas, G. Roussos, AI in cybersecurity: the paradox. IEEE Trans. Technol. Soc. **4**(2), 104–109 (2023)
67. M. Christen, B. Gordijn, M. Loi, *The ethics of cybersecurity*. Springer Nature (2020)
68. K. Macnish, J. Van der Ham, Ethics in cybersecurity research and practice. Technol. Soc. **63**, 101382 (2020)
69. D. Shou, Ethical considerations of sharing data for cybersecurity research, in *Financial Cryptography and Data Security: FC 2011 Workshops, RLCPS and WECSR 2011, Rodney Bay, St. Lucia, February 28-March 4, 2011, Revised Selected Papers 15*, pp. 169–177 (2012)