Mitigating Human Errors and Cognitive Bias for Human-AI Synergy in Cybersecurity

Federico Maria Cau[1], Lucio Davide Spano[1]

[1]University of Cagliari, Dept. of Mathematics and Computer Science, Via Ospedale 72, 09124, Cagliari, Italy

Abstract

Cybersecurity advancements necessitate effective measures to combat rising and sophisticated threats. Artificial Intelligence (AI) and eXplainable AI (XAI) solutions have demonstrated significant capabilities in predicting and responding to cyber threats. Moreover, integrating AI components with Intelligent User Interfaces (IUI) has been explored as a promising approach, emphasizing user experience and interaction policies. Despite these advancements, the primary challenge remains addressing human errors, particularly those induced by cognitive biases. This paper provides an overview of possible recommendations on AI integration with cybersecurity systems and human cognitive bias mitigation solutions.

Keywords

Cybersecurity, Artificial Intelligence, Intelligent User Interfaces, Cognitive biases

1. Introduction

With the advance of technology in different sectors, cybersecurity becomes crucial to protect against attacks and ensure digital safety [1]. As the complexity and frequency of cyber attacks rise, there is a need to employ different measures for identifying and countering emerging threats. Artificial Intelligence (AI) cybersecurity solutions [2, 3, 4, 5] have proven to be a good ally in fighting cybercrime, which, alongside eXplainable AI (XAI) techniques, have immense power in reducing and predicting cyber threats [6, 7]. Recent studies discussed the benefits of integrating AI components with an appropriate Intelligent User Interface (IUI), providing principles to apply when developing intelligent threat modelling tools, especially considering users' interaction guidelines and User Experience (UX) [8, 9, 10, 11, 12]. However, one of the main challenges that may invalidate the effectiveness of cybersecurity systems lies in human errors, often caused by users' cognitive biases [13, 14]. This calls for defining strategies to detect and mitigate human irrational judgements (e.g., optimism bias) [14, 15].

In this paper, we collect a set of shared problems and overcome strategies arising from the converging literature in cybersecurity considering Artificial Intelligence (AI), Intelligent User Interfaces (IUI), and human cognitive biases in decision-making. We start by outlining the emerging challenges of AI applications in cybersecurity systems and the necessity for

First International Workshop on Detection And Mitigation Of Cyber attacks that exploit human vuLnerabilitiES (DAMOCLES), colocated with AVI 2024, Arenzano (Genoa), Italy, June 3rd 2024

federicom.cau@unica.it (F. M. Cau); davide.spano@unica.it (L. D. Spano)
0000-0002-8261-3200 (F. M. Cau); 0000-0001-7106-0463 (L. D. Spano)

explainability resolutions. Next, we list the most encountered cognitive biases and phenomena that undermine users' decision-making and lead to potential errors, accompanied by mitigation approaches.

2. Related Work

In this section, we briefly summarize the emerging gaps and needs in the literature encompassing the fields of Artificial Intelligence (AI), Intelligent User Interface (IUI), and human cognitive bias in the cybersecurity domain.

2.1. Artificial Intelligence in Cybersecurity

Cybersecurity is continuously changing with the development of new technologies and the emergence of new threats [1]. The integration of Artificial Intelligence (AI) has the potential to significantly enhance cybersecurity systems by enabling them to identify and counter novel and unknown threats [16, 17]. Common approaches encompass AI tactics for identifying and surveilling malicious activities, detecting cyber threats, and safeguarding an organization's networks, which may include Expert Systems, Intelligent Agents, Deep Learning, and Reinforcement Learning [3, 4, 5, 18, 19]. Some practical examples of these techniques involve AI systems for Intrusion Detection [2, 20], Botnet Attack Detection [21, 22, 23, 24], Malware detection, analysis, and mitigation [25]. However, integrating complex and black-box AI systems undermines the transparency of these systems' decision-making processes. Adopting eXplainable AI (XAI) techniques becomes a starting point for providing insights into the rationale behind AI-driven decisions and enhancing overall transparency in the cybersecurity domain [6, 7, 26]. In addition to adopting Explainable AI (XAI) techniques, assessing AI confidence and robustness becomes crucial to avoid unintended behaviours of AI-based cybersecurity systems. Specifically, estimating measures like uncertainty and implausibility [27, 28] allows practitioners to make more informed decisions and adapt cybersecurity measures in responding to new threats.

2.2. Intelligent User Interfaces in Cybersecurity

Integrating Artificial Intelligence (AI) solutions in the cybersecurity ecosystem profoundly influences user interface (UI) design due to its ability to enhance user experiences and enable personalized interactions. This process is commonly referred to as Intelligent User Interface (IUI) design, ultimately leading to user interfaces that are user-centric, engaging, and effective in meeting user needs and expectations [8]. Previous literature already proposed potential designs of intelligent user interfaces for defending against Malicious Bot Attacks [24] and preventing Phishing Attacks [29, 30]. Instead, other research explored how users make decisions when engaging with these interfaces during phishing attacks [12], showing that the interface design was understandable and familiar to users. However, they posit a need for future research to determine the most effective malicious features to display and how to enhance users' interest and trust. Additionally, further research studied how cognitive bias affects users' decisions in a phishing detection scenario [11], revealing that the higher occurrence of hyperbolic discounting bias (i.e., choose immediate rewards over rewards that come later in the future) made it more easily identifiable by humans, reducing its effectiveness in deceiving participants. Conversely, the lower occurrence of authority bias (i.e., the tendency to be more influenced by the opinions and judgments of authority figures) proved more effective in phishing human participants. We will deepen the topic of cognitive biases in the next section.

2.3. Human Behavioral Decision-Making in Cybersecurity

A major challenge in the cybersecurity domain concerns the detection and mitigation of human cognitive biases, as they can influence decision-making, leading users to think irrationally in certain situations and make unreasonable judgments [13, 31]. Recent research gathered which are the most common cognitive biases based on different scenarios: for example, Gutzwiller et al. [32] found phenomena like confirmation bias, anchoring bias, and take-the-best heuristic are the most common among red teamers attackers. Furthermore, the authors of [33] investigated the role of four cognitive biases (i.e., selective perception, exposure to limited alternatives, adjustment and anchoring, and illusion of control) in anticipating and responding to Distributed-Denial-of-Service (DDoS) attacks. They highlighted several practical implications for managers in dealing with the increasing threat of cyberattacks like raising awareness, developing clear step-by-step tested and documented defense procedures, and identifying organizational vulnerabilities. Majumdar et al. [34] carried out a systematic literature review collecting human-related components (e.g., confirmation bias, availability bias, and framing effect) and risky habits (e.g., sharing passwords, accidental insider threats, and lack of perseverance) that impact cybersecurity practices, also suggesting solutions to overcome them among which: security awareness training, phishing simulations, and incident response plan. Alnifie and Kim [15] studied another relevant bias called optimism bias, which can result in an inaccurate perception of risks, leading to subjective decisions that lack objectivity. To reduce this bias, they suggest that employees regularly follow instructions from security teams, adhere to cybersecurity policies, and recognize optimism bias at both individual and organizational levels.

A novel security paradigm that emerged from recent literature [14] is referred to as cognitive security, where the authors emphasize the vulnerabilities in human cognitive processes (e.g., perception, attention, memory, and mental operations) that can be exploited by cognitive attacks, affecting performance and decision-making. The authors present several cognitive and technical defense methods to deter the kill chain (i.e., the stages of a cyberattack) of cognitive attacks, such as real-time tracking of cognitive attacks, identification of abnormal patterns in human behaviors, introducing compensation mechanisms to mitigate the impact of cognitive attacks, or reducing cognitive load during security incidents.

Another significant trend investigating human vulnerabilities involves cybersecurity games [35, 36], where participants make strategic decisions in a simulated environment and tackle real-world cyber threats to enhance their practical understanding of cybersecurity. Jalali et al. [37] developed a simulation game to assess decision-makers effectiveness in addressing two challenges in cybersecurity capability development: potential delays and uncertainties in predicting cyber incidents. They found that (i) decision-makers respond poorly to time delays in dynamic settings under uncertainty, and (ii) experienced managers did not perform better than inexperienced individuals in making proactive decisions about building cybersecurity capabilities. These results call for a strong need for training tools to underscore the drawbacks of a solely cognition-focused strategy and to grasp the impacts of feedback delays.

3. Discussion

The review of the literature work we discussed in Section 2 shows that the integration of AI is increasingly applied to defense mechanisms against cyber threats. From an interaction point of view, solutions that team together humans and AI-based agents are particularly relevant, especially for the tasks where AI-based solutions perform better than humans [1, 17]. This suggests that the synergistic teaming of humans and AI is a promising way to address cyber threats' dynamic and complex nature. However, the effectiveness of AI solutions in cybersecurity still poses several key challenges. One prominent issue is the shortage of skilled cybersecurity professionals [1, 16, 17]. This scarcity is a barrier even to the spread of AI-based solutions, which require professionals at least to set them up. The lack of professionals also sets a challenge to create and promote adequate educational programmes [3], which require expert human trainers again. However, building such programs is crucial for widespread awareness about cyber threats and fostering a culture of cybersecurity consciousness among organizations and individuals.

Another critical consideration is the design of computing platforms resilient to AI-based adversarial threats [19]. Rather than treating security measures as an afterthought, there is a growing recognition of embedding resilience into computing infrastructure from the outset. This proactive approach reflects a shared responsibility among stakeholders to mitigate cybersecurity risks effectively.

Transparency and interpretability are fundamental principles in deploying AI-driven cybersecurity solutions [16]. Biased data and decision-making processes pose significant challenges, as the opaque nature of AI models complicates understanding their logic and outcomes. It is often difficult for an administrator to understand the AI system logic in the event of a security breach. AI systems sometimes provide inaccurate findings in the form of false positives, which mislead security experts, jeopardizing the entire system's integrity [17]. Assimilating eXplainable AI solutions [6, 7] along with factors such as AI robustness and uncertainty [27, 28] is essential for maintaining trust and confidence in cybersecurity systems. Additionally, it is worth mentioning that XAI techniques can face security attacks, which emphasizes the need to carry out experimental studies of the impacts of various attacks on XAI methodologies, together with a balance between the security and usability of XAI-integrated cybersecurity systems [26].

Moreover, cognitive biases inherent in human decision-making introduce additional complexities to cybersecurity strategies [32]. The presence of biases, such as the take-the-best heuristic, confirmation bias, optimism bias, and anchoring bias [34], along with other phenomena like framing effects, sunk cost, irrational escalation, and the illusion of control [33], poses challenges for measurement. Future research should focus on inducing and assessing the exhibition of biased behaviour, moving away from over-reliance on observational assessment. Additionally, researchers should develop experimental designs and measures specifically designed to elicit particular biases.

In particular, optimism bias [15] refers to the tendency of individuals, regardless of their capacity, to perceive risks inappropriately. They often believe they are not vulnerable or exhibit overconfidence in the effectiveness of security measures: essentially, they think, "I/we won't be a target." To address this bias, researchers can explore longitudinal studies that track the development and evolution of optimism bias over time. Additionally, evaluating the effectiveness of different interventions, such as training programs, awareness campaigns, and educational initiatives, can help mitigate this bias. Furthermore, considering AI approaches may provide valuable insights into managing optimism bias.

Unique conditions significantly impact which biases are possible to study. For instance, it would be tough to investigate illusory correlation in a context lacking relevant data for correlation. Similarly, studying sunk cost [32] would be challenging if no resources were utilized. Some biases remain understudied, including the availability heuristic, default effect, and information-pooling bias. Additionally, social engineering techniques exploit cognitive biases to manipulate user behaviour [14], highlighting the importance of user training and awareness programs. Cultivating critical thinking skills and promoting a culture of cybersecurity consciousness is essential for defending against cognitive attacks and enhancing overall cybersecurity resilience.

4. Conclusion and Future Work

This paper summarised the shared needs and shortcomings of Artificial Intelligence (AI) solutions and human decision-making biases in cybersecurity. We discussed common points that emerge from the literature and provide potential directions against cybersecurity threats, which we outline as follows. The first significant aspect regards the importance of human-AI collaboration, urging the promotion of suitable professional programs to address the shortage of skilled cybersecurity experts. The second point highlights the necessity for transparency and explainability in cybersecurity AI solutions, revealing the necessity of planning new procedures to defend against AI-based cybersecurity attacks. Finally, the last trait calls for training programs to detect and measure cognitive biases, along with experimental settings that stimulate these biases, intending to elicit users' awareness and foster the growth of the cybersecurity culture.

Follow-up studies need to consider these aspects for a better AI-based cybersecurity systems administration, ensuring effective cyber-threat detection and mitigation by appropriately addressing human cognitive biases.

Acknowledgments

This work has been supported by the Italian Ministry of University and Research (MUR) and by the European Union - NextGenerationEU, under grant PRIN 2022 PNRR "DAMOCLES: Detection And Mitigation Of Cyber attacks that exploit human vuLnerabilitiES" (Grant P2022FXP5B).

References

[1] W. S. Admass, Y. Y. Munaye, A. A. Diro, Cyber security: State of the art, challenges and future directions, Cyber Security and Applications 2 (2024) 100031. URL: https://www.sciencedirect.com/science/article/pii/S2772918423000188. doi:https://doi.org/10.1016/j.csa.2023.100031.

[2] S. Alzughaibi, S. El Khediri, A cloud intrusion detection systems based on dnn using backpropagation and pso on the cse-cic-ids2018 dataset, Applied Sciences 13 (2023). URL: https://www.mdpi.com/2076-3417/13/4/2276. doi:10.3390/app13042276.

[3] A. Chakraborty, A. Biswas, A. Khan, Artificial Intelligence for Cybersecurity: Threats, Attacks and Mitigation, 2023, pp. 3–25. doi:10.1007/978-3-031-12419-8_1.

[4] M. Corbett, S. Sajal, Ai in cybersecurity, in: 2023 Intermountain Engineering, Technology and Computing (IETC), 2023, pp. 334–338. doi:10.1109/IETC57902.2023.10152034.

[5] M. Rizvi, Enhancing cybersecurity: The power of artificial intelligence in threat detection and prevention, International Journal of Advanced Engineering Research and Science 10 (2023). URL: https://journal-repository.theshillonga.com/index.php/ijaers/article/view/6352.

[6] F. Charmet, H. C. Tanuwidjaja, S. Ayoubi, P.-F. Gimenez, Y. Han, H. Jmila, G. Blanc, T. Takahashi, Z. Zhang, Explainable artificial intelligence for cybersecurity: a literature survey, Annals of Telecommunications 77 (2022) 789–812. URL: https://doi.org/10.1007/s12243-022-00926-7. doi:10.1007/s12243-022-00926-7.

[7] G. Rjoub, J. Bentahar, O. Abdel Wahab, R. Mizouni, A. Song, R. Cohen, H. Otrok, A. Mourad, A survey on explainable artificial intelligence for cybersecurity, IEEE Transactions on Network and Service Management 20 (2023) 5115–5140. doi:10.1109/TNSM.2023.3282740.

[8] J. E. T. Akinsola, S. I. Akinseinde, O. Kalesanwo, M. A. Adeagbo, K. Oladapo, A. A. Awoseyi, F. A. Kasali, Application of artificial intelligence in user interfaces design for cyber security threat modeling, Intelligent User Interfaces [Working Title] (2021). URL: https://api.semanticscholar.org/CorpusID:233567804.

[9] A. Jayatilaka, N. A. G. Arachchilage, M. A. Babar, Why people still fall for phishing emails: An empirical investigation into how users make email response decisions, ArXiv abs/2401.13199 (2024). URL: https://api.semanticscholar.org/CorpusID:267200262.

[10] S. Zheng, I. Becker, Presenting suspicious details in User-Facing e-mail headers does not improve phishing detection, in: Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022), USENIX Association, Boston, MA, 2022, pp. 253–271. URL: https://www.usenix.org/conference/soups2022/presentation/zheng.

[11] M. Sharma, M. Kumar, C. Gonzalez, V. Dutt, How the presence of cognitive biases in phishing emails affects human decision-making?, in: M. Tanveer, S. Agarwal, S. Ozawa, A. Ekbal, A. Jatowt (Eds.), Neural Information Processing, Springer Nature Singapore, Singapore, 2023, pp. 550–560.

[12] G. Desolda, J. Aneke, C. Ardito, R. Lanzilotti, M. F. Costabile, Explanations in warning dialogs to help users defend against phishing attacks, International Journal of Human-Computer Studies 176 (2023) 103056. URL: https://www.sciencedirect.com/science/article/pii/S1071581923000654. doi:https://doi.org/10.1016/j.ijhcs.2023.103056.

[13] T. Albalawi, K. Ghazinour, A. Melton, Quantifying the effect of cognitive bias on security decision-making for authentication methods, in: J. Ren, A. Hussain, H. Zhao, K. Huang, J. Zheng, J. Cai, R. Chen, Y. Xiao (Eds.), Advances in Brain Inspired Cognitive Systems, Springer International Publishing, Cham, 2020, pp. 139–154.

[14] L. Huang, Q. Zhu, An introduction of system-scientific approaches to cognitive security, ArXiv abs/2301.05920 (2023). URL: https://api.semanticscholar.org/CorpusID:255942622.

[15] K. Alnifie, C. Kim, Appraising the manifestation of optimism bias and its impact on human

perception of cyber security: A meta analysis, Journal of Information Security 14 (2023) 93–110. doi:10.4236/jis.2023.142007.

[16] R. R. Shanthi, N. K. Sasi, P. Gouthaman, A new era of cybersecurity: The influence of artificial intelligence, in: 2023 International Conference on Networking and Communications (ICNWC), 2023, pp. 1–4. doi:10.1109/ICNWC57852.2023.10127453.

[17] N. Mohamed, Current trends in ai and ml for cybersecurity: A state-of-the-art survey, Cogent Engineering 10 (2023) 2272358. URL: https://doi.org/10.1080/23311916.2023.2272358. doi:10.1080/23311916.2023.2272358. arXiv:https://doi.org/10.1080/23311916.2023.2272358.

[18] M. Lourens, A. P. Dabral, D. Gangodkar, N. Rathour, C. N. Tida, A. Chadha, Integration of ai with the cybersecurity: A detailed systematic review with the practical issues and challenges, in: 2022 5th International Conference on Contemporary Computing and Informatics (IC3I), 2022, pp. 1290–1295. doi:10.1109/IC3I56241.2022.10073040.

[19] E. Adi, Z. Baig, S. Zeadally, Artificial intelligence for cybersecurity: Offensive tactics, mitigation techniques and future directions, Applied Cybersecurity & Internet Governance 1 (2022) 2–24.

[20] S. V. N. Santhosh Kumar, M. Selvi, A. Kannan, A. D. Doulamis, A comprehensive survey on machine learning-based intrusion detection systems for secure communication in internet of things, Intell. Neuroscience 2023 (2023). URL: https://doi.org/10.1155/2023/8981988. doi:10.1155/2023/8981988.

[21] P. Dixit, R. Kohli, A. Acevedo-Duque, R. R. Gonzalez-Diaz, R. H. Jhaveri, Comparing and analyzing applications of intelligent techniques in cyberattack detection, Security and Communication Networks 2021 (2021) 5561816. URL: https://doi.org/10.1155/2021/5561816. doi:10.1155/2021/5561816.

[22] S. S., D. G., D. S., An intelligent iot attack detection framework using effective edge ai based computing, Indian Journal of Computer Science and Engineering 13 (2022) 1156–1167. doi:10.21817/indjcse/2022/v13i4/221304059.

[23] S. Afrifa, V. Varadarajan, P. Appiahene, T. Zhang, E. A. Domfeh, Ensemble machine learning techniques for accurate and efficient detection of botnet attacks in connected computers, Eng 4 (2023) 650–664. URL: https://www.mdpi.com/2673-4117/4/1/39. doi:10.3390/eng4010039.

[24] C. Dinuwan, H. Amandakoon, I. Aberathne, T. Wimalarathna, R. Ratnayake, Ai-powered detection and prevention tool to secure apis from malicious bot attacks, in: T. Senjyu, C. So-In, A. Joshi (Eds.), Smart Trends in Computing and Communications, Springer Nature Singapore, Singapore, 2023, pp. 555–566.

[25] A. Djenna, A. Bouridane, S. Rubab, I. M. Marou, Artificial Intelligence-Based Malware Detection, Analysis, and Mitigation, Symmetry 15 (2023) 677. doi:10.3390/sym15030677.

[26] G. Srivastava, R. H. Jhaveri, S. Bhattacharya, S. Pandya, Rajeswari, P. K. R. Maddikunta, G. Yenduri, J. G. Hall, M. Alazab, T. R. Gadekallu, Xai for cybersecurity: State of the art, challenges, open issues and future directions, 2022. arXiv:2206.03585.

[27] J. DeMarchi, R. Rijken, J. Melrose, B. Madahar, G. Fumera, F. Roli, E. Ledda, M. Aktaş, F. Kurth, P. Baggenstoss, B. Pelzer, L. Kanestad, Evaluation of robustness metrics for defense of machine learning systems, 2023, pp. 1–12. doi:10.1109/ICMCIS59922.2023.10253593.

[28] F. O. Catak, T. Yue, S. Ali, Uncertainty-aware prediction validator in deep learning models for cyber-physical system data, ACM Trans. Softw. Eng. Methodol. 31 (2022). URL: https://doi.org/10.1145/3527451. doi:10.1145/3527451.

[29] J. Aneke, C. Ardito, G. Desolda, Towards intelligent user interfaces to prevent phishing attacks, 2020, pp. 279–288. doi:10.18573/book3.ak.

[30] J. Aneke, C. Ardito, G. Desolda, Designing an Intelligent User Interface for Preventing Phishing Attacks, in: J. A. Nocera, A. Parmaxi, M. Winckler, F. Loizides, C. Ardito, G. Bhutkar, P. Dannenmann (Eds.), 17th IFIP Conference on Human-Computer Interaction (INTERACT), volume LNCS-11930 of *Beyond Interactions*, Springer International Publishing, Paphos, Cyprus, 2019, pp. 97–106. URL: https://inria.hal.science/hal-03188818. doi:10.1007/978-3-030-46540-7_10, part 3: Workshop on Handling Security, Usability, User Experience and Reliability in User-Centered Development Processes.

[31] S. Hai-Jew, The Electronic Hive Mind and Cybersecurity: Mass-Scale Human Cognitive Limits to Explain the "Weakest Link" in Cybersecurity, 2018, pp. 206 – 262. doi:10.4018/978-1-5225-5927-6.ch011.

[32] R. S. Gutzwiller, K. J. Ferguson-Walter, S. J. Fugate, Are cyber attackers thinking fast and slow? exploratory analysis reveals evidence of decision-making biases in red teamers, Proceedings of the Human Factors and Ergonomics Society Annual Meeting 63 (2019) 427–431. URL: https://doi.org/10.1177/1071181319631096. doi:10.1177/1071181319631096. arXiv:https://doi.org/10.1177/1071181319631096.

[33] A. Ceric, P. Holland, The role of cognitive biases in anticipating and responding to cyberattacks, Information Technology & People 32 (2019) 171–188. URL: https://doi.org/10.1108/ITP-11-2017-0390. doi:10.1108/ITP-11-2017-0390.

[34] N. Majumdar, V. Ramteke, Human elements impacting risky habits in cybersecurity, volume 2519, 2022, p. 030006. doi:10.1063/5.0110624.

[35] M. Abdallah, D. Woods, P. N. Ardabili, I. M. Khalil, T. N. Cason, S. Sundaram, S. Bagchi, Bascps: How does behavioral decision making impact the security of cyber-physical systems?, ArXiv abs/2004.01958 (2020). URL: https://api.semanticscholar.org/CorpusID:214802678.

[36] T. Malloy, C. Gonzalez, Learning to defend by attacking (and vice-versa): Transfer of learning in cybersecurity games, in: 2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&amp;PW), IEEE Computer Society, Los Alamitos, CA, USA, 2023, pp. 458–464. URL: https://doi.ieeecomputersociety.org/10.1109/EuroSPW59978.2023.00056. doi:10.1109/EuroSPW59978.2023.00056.

[37] M. S. Jalali, M. D. Siegel, S. E. Madnick, Decision-making and biases in cybersecurity capability development: Evidence from a simulation game experiment, Cyberspace Law eJournal (2017). URL: https://api.semanticscholar.org/CorpusID:7219775.