Interdisciplinary Research Paper

Nick Dorsey

Old Dominion University

Dr. Kat LaFever

03 Aug 2025

Introduction

In a digital world, businesses must make sure that their systems are well-insulated against more advanced cybersecurity threats. One method of system security testing is a penetration test; ethical hackers can model authentic attacks. As AI grows more common, a stronger urge to have AI aid or support ethical hackers has been exerted. This technological leap presents a conundrum: To what extent should artificial intelligence be allowed to augment ethical hackers performing penetration testing on organizations? To critically understand the implications of this question, this paper relies on computer science, ethics, and psychology. Computer science plays a significant role in AI capabilities and limitations. Ethics provides a means to examine the ethical implications of entrusting machines to interfere with sensitive security procedures. Psychology might help us to comprehend the human reactions to AI assistance, especially related to trust and decision-making. Both disciplines bring distinctive and essential answers to this complicated question. An interdisciplinary approach facilitates a more holistic understanding through the integration of various methodologies and frameworks. It is an approach required because the issue of the role of AI in penetration testing is not only technical but also ethical and psychological. This study is directly related to my major in cybersecurity and my career aspirations in ethical hacking and systems analysis. With the growing use of AI tools in the security sector, practitioners should be able to evaluate the potential and ethical limits of this technology critically. Including this paper in my IDS 493 portfolio shows that I can address multifaceted, complex issues relevant to my future career.

Definition of Key Terms

Key terms used in this paper include artificial intelligence, ethical hacking, penetration testing, cognitive trust, and autonomous decision-making, which should be explained.

According to Biplob and Rahman (2025), artificial intelligence is a set of algorithms and systems that exhibit goal-directed behavior and decision-making abilities traditionally attributed to human intelligence. In cybersecurity, AI encompasses applications such as machine learning models that detect and exploit vulnerabilities in systems.

Baxromov (2025) defines ethical hacking as the legitimate effort to obtain unauthorized access to a computer system, application, or data to detect security vulnerabilities that malicious hackers can exploit. Ethical hackers work with the organization's approval to improve security.

According to Floridi et al. (2018), penetration testing is a process of assessing the security of a computer system or network by mimicking an attack by a malicious actor.

Cognitive trust is defined as the rational belief of one person in the reliability and competence of another person (Devlin et al., 2025).

Lastly, autonomous decision-making in AI is described by Terziyan et al. (2025) as the ability of a system to make decisions and perform tasks without direct human involvement, based on embedded logic or trained behavior.

These definitions help to understand the scope of the paper and establish a common understanding of the key terms used throughout the discussion.

Computer Science

In computer science terms, artificial intelligence is already being utilized to augment penetration testing by improving speed, efficiency, and coverage. According to Biplob and Rahman (2025), AI tools such as machine learning classifiers can identify known vulnerabilities

within a fraction of the time it would take a human. AI-powered tools like fuzzers and network scanners have already been deployed in large organizations. Ethical hackers can use IA to simulate the behavior of attackers based on data gathered during actual cyberattacks, providing them with high-fidelity simulations and predictive capabilities (Zhang et al., 2025). Moreover, reinforcement learning algorithms are trained to emulate advanced persistent threats (APTs). These models are capable of independently navigating networks, raising privileges, and locating vulnerabilities, which would take human ethical hackers much longer (Ruohonen & Saddiqa, 2025). These features assist testers in prioritizing the vulnerabilities to be addressed first, depending on risk profiles, thus streamlining defensive measures.

Nevertheless, AI has its limits. AI systems are unable to recognize context-specific threats and cannot identify zero-day vulnerabilities that have not been identified in security databases. Additionally, AI systems are susceptible to adversarial attacks or false positives, where genuine actions are incorrectly identified as malicious (Zhang et al., 2025). Although these tools offer massive computing capabilities, they are best used to complement rather than substitute human judgment and creativity in penetration testing.

Ethics

The ethical issues of AI penetration testing include accountability, moral agency, and the possibility of harm. As Thapaliya and Dhital (2025) point out, delegating the decision-making process to machines blurs the lines of responsibility: who is to blame in case of a system crash or access to unauthorized data access, the programmer, the user, or the AI? In security situations where trust and legality are paramount, ambiguous accountability poses significant risks. Another ethical issue is autonomy. Bahangulu and Berko (2025) assert that machines lack the moral reasoning to understand the consequences of their actions. They are thus not qualified to

undertake duties that involve judgment of legal boundaries. Consider a case where an AI port scan is legal in one jurisdiction and illegal in another. When not controlled by humans, the risk of violating ethical norms or regulations increases. Floridi et al. (2018) emphasize the importance of transparency and fairness in AI systems. The security advantage is in favor of companies that can afford expensive AI tools, so small businesses and non-profits are at a disadvantage. That poses ethical questions about equity and fairness in cybersecurity defense. In conclusion, the moral perspective proposes a more cautious, responsible application of AI to penetration testing, where the issues of human control, transparency, and justice in the digital realm are more prominent.

Psychology

Psychology offers valuable insights into the interaction between AI and humans during penetration testing by highlighting trust, decision-making, and human factors. Here, cognitive trust plays an important role. Romeo & Conti (2025) suggest that human trust in automation must be calibrated. Over-trust may result in automation complacency, where users accept AI outputs without question. A lack of trust leads to disuse, where valuable insights are dismissed. As Endsley (2018) explains, the cognitive burden on ethical hackers will rise or fall depending on the ease with which the AI tools integrate into their workflow. AI tools that offer vague or unclear recommendations generate stress and lower performance. Conversely, transparent and intuitive tools minimize cognitive load and enable humans to concentrate on strategic decision-making. The feeling of agency, the psychological sense of control over your actions, is also at risk. According to Rebera et al. (2025), delegating decisions to AI can leave humans feeling disconnected or disempowered, particularly when they object to AI-generated actions but lack the confidence to override them. This is especially risky in cybersecurity, where professional

responsibility is paramount. Integrating psychology can help to prevent AI integration from undermining user autonomy, decreasing engagement, or creating unhealthy levels of over-reliance on automation.

Common Ground

This interdisciplinary study reveals three significant findings that create a solid basis of shared ground between the fields of computer science, ethics, and psychology. To begin with, all three fields concur that artificial intelligence (AI) must be employed as an assistive technology, rather than a replacement for human ethical hackers. Whereas computer science highlights the immense computational capabilities, scalability, and pattern recognition abilities of AI, ethics and psychology warn against the loss of the human factor, the importance of personal responsibility, intuition, and situational awareness in cybersecurity operations. Second, the disciplines agree that human oversight is required. AI tools must operate under human control to maintain accountability and flexibility. Ethical hackers should have the right to interpret, override, or reject AI-generated outputs where needed. Without this, the risk of automation bias or moral disengagement rises. The human-in-the-loop model is crucial to the responsible and practical application of AI technologies (Romeo & Conti, 2025). Third, everyone agrees that contextual flexibility is essential. AI systems should be customized to suit various organizational objectives, regulatory environments, and psychological contexts (Endsley, 2018). There is no single AI framework that can adequately respond to the dynamic and sensitive nature of cybersecurity threats.

Disciplinary Conflicts

Although there are areas of convergence, there are also significant disagreements between the fields of computer science, ethics, and psychology. The most notable is the conflict

between automation and responsibility. Computer science-wise, the focus is on speed, scalability, and efficiency via AI-powered automation. Automation is regarded as a means to minimize human error and enhance consistency in penetration testing. Nevertheless, ethicists express concerns over ceding too much control to machines that lack moral judgment. They claim that responsibility and accountability should be left to human agents, particularly in cases where AI decisions may affect privacy, legality, or organizational reputation (Thapaliya & Dhital, 2025). A second significant point of contention concerns the conceptualization of trust. Psychology views trust as a subjective, dynamic process influenced by past experiences, cognitive biases, and perceived control. In computer science, trust is often reduced to quantifiable aspects such as system reliability or algorithm accuracy. This gap creates a scenario in which a technically accurate AI result may still be distrusted by human operators who feel unsure or excluded in the decision-making process (Romeo & Conti, 2025). To address these disciplinary frictions, developers ought to create explainable and transparent AI systems that integrate ethical theories and allow human understanding. This habit promotes accountability and trust.

Ch. 12: Constructing a More Comprehensive Understanding or Theory

The synthesis of computer science, ethics, and psychology offers a more detailed image of the role of artificial intelligence in ethical hacking. This paper advances a Collaborative AI-Augmented Ethical Hacking theory that integrates these views. The proposed model would involve using AI tools that effectively supplement, rather than substitute, human ethical hackers to guarantee that technological advances would not undermine moral responsibility and psychological integrity. The model suggests a tiered strategy of AI involvement where the level of AI autonomy is contingent upon the riskiness and sensitivity of the testing environment. AI

may be used semi-autonomously in low-threat environments, such as routine vulnerability scans or password strength tests, to improve efficiency (Thapaliya & Dhital, 2025). However, in critical areas like testing hospital networks or financial systems, significant human control is required to prevent unintended injury or legal violations. Additionally, the theory proposes a trust calibration mechanism, meaning that efficient use of AI necessitates system transparency and ongoing user training. Ethical principles must be incorporated into AI algorithms, and users should be empowered to interpret and challenge AI outputs. This unified system protects against excessive dependence and alienation, so that AI becomes a cooperative tool, maintaining technical efficiency and moral integrity.

Ch. 13: Reflecting On, Testing, and Communicating the Understanding or Theory

Reflecting on the theory of Collaborative AI-Augmented Ethical Hacking usually demands for a tactical evaluation of how it is implemented, tested, and improved in the practice of cybersecurity. A good way of testing the model is by using pilot penetration testing programs that involve human ethical hackers and AI systems. Such controlled experiments can measure performance measures like detection rates, false positives, response times, and ethical decision outcomes (Bahangulu & Berko, 2025). Moreover, the quality of team trust and human-AI interaction can be assessed with the help of validated psychological tests to establish the extent to which ethical hackers can cooperate with AI systems under different circumstances (Devlin et al., 2025). To make the model robust, it must be applied in various sectors, including those where compliance is strict, like the healthcare, banking, and government sectors. Such environments have a high legal and ethical cost of AI errors, which makes them a perfect place to test the theory and its limitations. The practitioner feedback through interviews, focus groups, and observational studies can be vital in refining the model. It is also important to communicate

the theory effectively. This includes publishing interdisciplinary research in cybersecurity, psychology, and ethics journals, and conducting workshops and webinars with practitioners. Policymakers and executives (the non-technical stakeholders) should be addressed using plain, non-jargon-filled language designed to facilitate widespread understanding and responsible action.

Conclusion

The question of whether to employ an artificial intelligence to augment ethical hackers in penetration testing is something that cannot be answered under a particular disciplinary lens. As the present study demonstrates, such an interdisciplinary approach that would include computer science, ethics, and psychology is needed to grasp the fine details of introducing AI into the practice of cybersecurity in all its richness. Despite the unequalled benefits in terms of speed, scalability, and data analysis, the implementation of AI raises significant doubts regarding accountability, transparency, and human trust. The two areas are equally important, and when combined, they form a more balanced and practical image. Computer science highlights the benefits of AI in automating repetitive procedures, as well as in modeling sophisticated attack trends and enabling ethical hackers to work more efficiently. Ethics reminds us that we have to ensure that the power of technology is weighed against moral responsibility, to ensure that the consequences of actions on people and systems are taken through human judgment. The importance of trust, agency, and user experience is indicated in the field of psychology and can be lost if AI is misapplied or misunderstood. In sum, these insights help to form the general conclusion that AI should be a team member, an assistant to human capabilities, not an equivalent. The Collaborative AI-Augmented Ethical Hacking theory is an ethical and practical point of light. Cyber threats have never been more prolific, and the adoption of such

interdisciplinary structures is critical to the development of secure, accountable, and resilient online environments.

References

Bahangulu, J. K., & Berko, L. O. (2025). Algorithmic bias, data ethics, and governance: Ensuring fairness, transparency, and compliance in AI-powered business analytics applications. *World Journal of Advanced Research and Reviews*, *25*(2), 1746-1763. https://eprint.scholarsrepository.com/id/eprint/847/

Baxromov, K. (2025). ETHICAL HACKERS: WHO THEY ARE AND WHY COMPANIES REWARD THEM. *Теоретические аспекты становления педагогических наук*, *4*(7), 129-132. https://inlibrary.uz/index.php/tafps/article/view/79448

Biplob, M. B., & Rahman, A. (2025). The Role of Explainable AI in Automated Software Testing: Opportunities and Challenges. https://www.preprints.org/frontend/manuscript/ac4c260521783ab5c71424fcc1bb4bfa/download_pub

Devlin, J. F., Roy, S. K., Sekhon, H., Moin, S. M. A., & Sahiner, M. (2025). Trust and FinTech: A review and research agenda. *Electronic Markets*, *35*(1), 62. https://link.springer.com/article/10.1007/s12525-025-00803-w

Thapaliya, S., & Dhital, S. (2025). AI-Augmented Penetration Testing: A New Frontier in Ethical Hacking. *International Journal of Atharva*, *3*(2), 28–37. https://nepjol.info/index.php/ija/article/view/80099

Endsley, M. R. (2018). Automation and situation awareness. In *Automation and human performance* (pp. 163–181). CRC Press. https://www.taylorfrancis.com/chapters/edit/10.1201/9781315137957-8/automation-situation-awareness-mica-endsley

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and machines*, *28*(4), 689-707. https://link.springer.com/article/10.1007/S11023-018-9482-5

Rebera, A. P., Lauwaert, L., & Oimann, A. K. (2025). Hidden Risks: Artificial Intelligence and Hermeneutic Harm. *Minds and Machines*, *35*(3), 33. https://link.springer.com/article/10.1007/s11023-025-09733-0

Romeo, G., & Conti, D. (2025). Exploring automation bias in human–AI collaboration: a review and implications for explainable AI. *AI & SOCIETY*, 1–20. https://link.springer.com/article/10.1007/s00146-025-02422-7

Ruohonen, J., & Saddiqa, M. (2025). What Do We Know About the Psychology of Insider Threats?. In *International Conference on Digital Forensics and Cyber Crime* (pp. 186-211). Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-031-89363-6_11

Terziyan, V., Tiihonen, T., Shukla, A. K., Gryshko, S., Golovianko, M., Terziyan, O., & Vitko, O. (2025). Towards ethical evolution: responsible autonomy of artificial intelligence across generations. *AI and Ethics*, 1-26. https://link.springer.com/article/10.1007/s43681-025-00759-9

Zhang, J., Bu, H., Wen, H., Liu, Y., Fei, H., Xi, R., ... & Meng, D. (2025). When LLMs meet cybersecurity: a systematic. https://cybersecurity.springeropen.com/counter/pdf/10.1186/s42400-025-00361-w.pdf