Ready for Battle?: Legal Considerations for Upcoming AI Hacker and Vulnerability Issues

Lee Jin-Myeong, Yoon Sang-Pil
School of Cybersecurity, Korea University, Seoul, South Korea
wjsp8392@korea.ac.kr, ssangbbi@gmail.com

Abstract

While the legal framework for AI is under much discussion, we need to consider the future where AI will be adopted into every aspect of our daily lives. As a software package, AI cannot be perfect and is greatly influenced by the intent of its user or its creator. In this regard, AI can be used to breach or strengthen cybersecurity. As such, there has been much discussions on how to make use of AI technology for cybersecurity while adopting it at the same time. In fact, there are already many cases of using AI to identify vulnerabilities or using AI for hacking. So what are the most reasonable set of regulatory measures to ensure balanced use of AI? This research seeks to identify the legal challenges of dealing with AI hacking. To accomplish this objective, we need to deduce elements of AI hacking from cybersecurity properties of AI technology and find real cases to analyze. Based on analysis of these real cases, this paper will propose the possible legal challenges and tasks ahead.

Background and Methodology

The fear of AI ruling over humankind is now an outdated story. Today, the discussions concerning AI are focused on how to make use of AI as a tool. And the EU has quickly developed a legal framework for AI(European Commission 2021). Areas such as biometric recognition and classification, public infrastructure, education and labor, law enforcement have been classified as High-risk AI, subject to strong legal enforcement. Any company that develops AI services in these areas must take the appropriate measures to ensure full legal compliance such as the implementation of risk management system, high quality data set, technical documentation, proper record-keeping, transparency, human oversight as well as measures for accuracy, robustness and cybersecurity. According to Article 15 of the said legal framework, these High-risk AI services must be resilient against exploits against the vulnerabilities of AI. To deal with such threats, the technical solutions that address AI specific vulnerabilities must include measures for data positioning attacks against data sets, input attacks that try to cause error in the model or flaws of the model itself.

This legal framework for AI which is the first in human history, only has a limited set of security regulations for the High-risk AI. However, similar to the regulations for the cyberspace, as AI services begin to proliferate, so will the regulations for them. AI is a set of software and is a dual-use tool that is greatly influenced by the intent of its user or its creator. AI tools developed and distributed by the private sector can be used by both the military powers or criminals(Sayler 2020). We need to prepare for the future when AI becomes prevalent and a deep part of our society. While it is not possible to fully control our future, we need to predict possible side effects and discuss safety against such issues(Taddeo and Floridi 2019).

This paper seeks to identify issues in regulations of AI based on the analysis of how AI can be used for hacking. AI can be used to breach or strengthen cybersecurity. Without consideration for security, the various tools already distributed can undermine security by becoming the source of vulnerabilities as finding vulnerabilities begin with complex and loosely connected systems as well as unnecessarily long codes. In the end, AI that can discover new vulnerabilities will become useful tool for governments, criminals and hackers alike(Schneier 2021). By analyzing this phenomenon and looking into possible regulations beforehand, this paper seeks to contribute to the development of better security policy against AI hacking.

To describe our methodology, the main goals of this paper include:

- Identify properties of AI from a cybersecurity perspective
- Identify the attracting factors for hacker from AI properties
- Find AI hacking cases and outlook for each of the attracting factor
- Deduce relevant legal issues and propose the legal challenges

To achieve the goals above, this paper will look into the possibility of AI hacking and real world cases based on the properties of AI first. Using these facts, we can apply the relevant legal principles to perform the norm judgement and identify legal issues. Lastly, systematic resolutions to resolve these legal issues will be proposed.

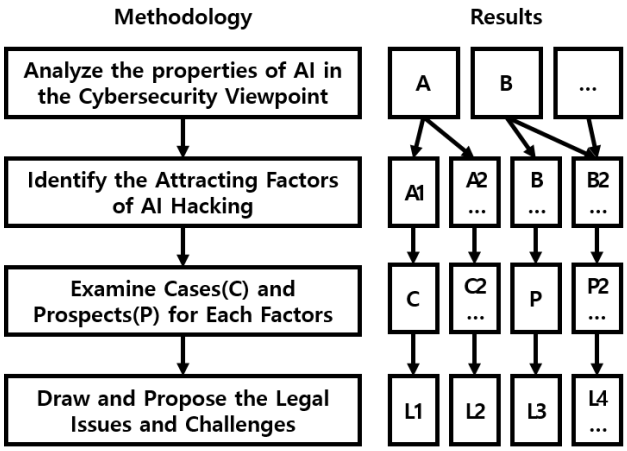

Figure 1: Methodology

Understanding AI Hacker

The Cybersecurity Properties of AI

Some of the properties unique to AI greatly influence the practice of cybersecurity. For example, according to analytical research in virtual hacker robots, characteristics of compatibility and seamless integration, robotics automation, continuous upgrade capability, and real-time have been discovered(Yuehong 2021). Other research on cases and outlook of abuse of AI have identified the traits of AI to be dual-use, efficiency, exceeding human capabilities, anonymity, and scalability(Brundage et al. 2018). In addition, some attackers are using AI hacking to make use of the advantages provided by AI such as overcoming the air gap barrier, encoding, strategy customization, processing rich data of target-specific intelligence, and scalability(Chung et al. 2021). Also, tools like AI malware can automatically interpret context and provide functions such as adaptiveness, autonomous decisions, and evading detection(Darktrace 2018).

The common elements of these functions point to compatibility, dual-use, scalability, invisibility, independency and creativity as the properties of AI. This allows hackers to perform very effective and efficient attacks compared to previous attacks. Hackers are able to make use of various platforms and tools using AI to effectively find vulnerabilities and attack without the need for direct commands or control. In some cases, AI provided ways to attack that even the hackers did not come up with.

Properties of AI	
Compatibility	Easy integration with intelligence platforms, compatibility with scanning tools
Dual-use	Can be used for both attack and defense
Scalability	Expanding target system by using abstract algorithm, evolve by learning from data from different systems and environment
Invisibility	Difficult to reverse engineer by learning models, increased anonymity, obfuscation
Independence	Able to make independent decisions
Creativity	Can find solutions not previously used before

Table 1: Security focused properties of AI

Identifying the Attracting Factors for Attackers from the Properties of AI

The kill-chain of the average hacking attacks as defined by Lockheed Martin and the benefits provided by AI hacking for each stage of the attack can be summarized as follows:

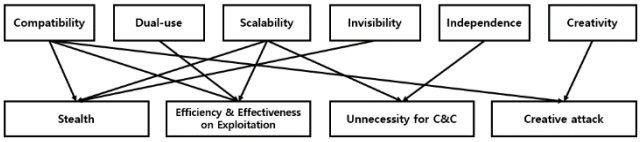

Figure 2: Benefits for the attacker by the properties of AI

First, the attacker can take advantage of stealth provided by better invisibility and scalability as well as compatibility. Therefore, at the reconnaissance stage, easier access can be achieved by circumventing the authentication process through imitating the normal state of a system based on stealthy information gathering. For example, an AI algorithm publicly available as open source can easily neutralize and disable the mechanism that identifies attackers based on signatures or TTPs. Authentication·trust modules can be circumvented by imitating normal system state based on information gathered on the target system. The benefit of such stealth comes at the Delivery stage where as we can see from the DeepLocker case that made use of the DNN AI model, it is possible to make it next to impossible to reverse engineer the attack unless specific trigger conditions are fulfilled (Ciancaglini et al. 2020).

At the Exploitation stage, AI properties of compatibility, dual-use and scalability increase the effectiveness and efficiency of vulnerability scanning. SAIVS, developed in January 2016, was able to discover vulnerabilities in web applications such as XSS and SQL injection vulnerabilities. (Takaesu and Terada, 2016). While the scanning above was done on simple web applications, the vulnerability scanners that have been developed recently can gather high quality information from data sets provided by OSINT and provide more accurate and effective information by gathering information for next hop targeting(Mirsky et al. 2021). Deep

learning based algorithms such as NeuFuzz, VulDeePecker and LAFuzz have been proven to be able to efficiently find vulnerabilities in codes without human intervention(Li et al. 2018; Wang et al. 2019; Wang et al. 2020).

The attacker can make use of the independence and scalability properties of AI to deliver commands during the C&C stage of the attack since AI can make autonomous decisions based on calculations and execute them. Therefore, the attacker does not need to send out the execution command. And even with no physical contact, once AI has been injected, AI will gather information of its system environment and autonomously determine the parameters and payload. Based on Adversarial DRL at opcode level, the ADVERSARIALuscator can obfuscate the metamorphic malware and produce several modified instances. This clearly shows that the malware can gather information within the target system and modify the internal structure most properly(Sewak et al. 2021).

Furthermore, based on the creativity and compatibility properties, AI can execute attacks in new ways that human beings cannot think of, as seen from example cases of DeepHack and DeepExpload announced at the DEFCON 2017. In these cases, with no other information than that provided by the target server responses, AI built a neural network capable of producing SQL injection text strings, thus automating the process of web based database hacking. Moreover, due to its compatibility property that allows the organic gathering in diverse data environment, the success rate of the attack was higher(Ciancaglini et al., 2020).

Case Analysis & Policy Implication

We can summarize the direction for regulation against hackers using AI by taking away an incentive in four ways: ① design transparency principle to reduce the benefit of stealth, ② undermine cost-effectiveness of attack life cycle, ③ retain minimum external control on AI beyond physical limitation, and ④ establish a response system for preliminary cybersecurity framework against unpredictability and enlarged attack space by a creative attack.

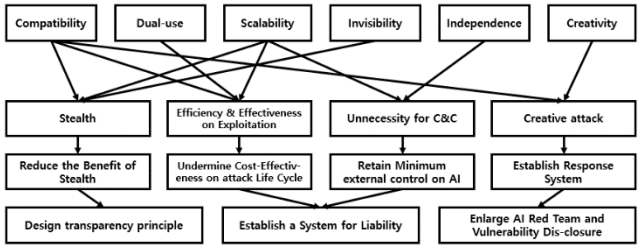

Figure 3: Legal Issues & Policy Implications to Regulate AI Hacking

Reduce the Benefit of Stealth : Design transparency principle

Legal enforcement of strong authentication and recording obligations can reduce the benefits of Stealth, an incentive for AI hackers. Therefore, a verification system is needed to implement security measures from the designing stage of AI development(Benjamins et al. 2019). This will act as a certification system to make so-called “Green AI”.

AI hacker can constantly make changes or adapt its attack patterns based on its adaptiveness or creativity properties. We need to be able to deal with intransparency or unpredictability issues that arise from such properties(Taddeo et al. 2019). To do so, a method to record all AI activities can be considered.

Such certification or mandatory record-keeping can act as an industrial regulation. However, considering the social impact of AI and the fact that AI is not just a simple game program, it should be provided in the most reliable form possible. In other words, AI regulations must be designed from a product safety perspective.

Undermine Cost-effectiveness and Retain Minimum External Control : Establish a System for Liability

Legal consequences or the liability for illegal hacking need to be strengthened to suppress hacking. In order to enforce legal mandates, we need to be able to resolve the liability issue. Liability requires a clear causal relationship between the action and the result of the said action to be proved. However in AI hacker case, it is tough to identify and prove specific facts pertaining to an attack. We need to adopt a strict liability standard to consider such characteristics(Gerstner 1993). Regarding the creator of AI hacker, the creator should be forced to take the liability for his creation whether or not requirements for the liability such as negligence or intent, action taken have been proven. This allows minimum control over AI by connecting software operating in a separate environment and the subjects utilizing it.

In addition, by applying tighter legal controls on ownership and sharing of dual-use items, developers who design and distributors who disclose AI to be used in attacks can be legally liable. This could increase the cost of using AI for attacks. This has already been discussed in related legislation in the bio-industry and WMD sectors of each country.

Establish Response System : Enlarge AI Red Team and Vulnerability Disclosure

From an organizational point of view, we need to build and operate an AI red team capable of carrying out penetration testing to gain access into our system in order to discover new vulnerabilities of our system(Schmidt et al. 2021). Spe-

cifically, when we can maximize diverse and creative methods of access, we can discover the most number of vulnerabilities and flaws in our security(Maillart et al. 2017). To achieve this goal, relevant groups must be able to research and publicize the newly identified vulnerabilities and be free from legal repercussions(McKinney 2008). The CISA of the U.S. requested various U.S. federal agencies to establish vulnerability disclosure policy through the Binding Operational Directive 20-01 announced in 2020. Currently, the said policy does not specifically include AI systems but should be extended to include the vulnerability disclosure policy for all new AI systems implemented (Dempsey and Grotto 2021).

Conclusion

With the evolution of data and IT technology, AI technology will also evolve and considering the characteristics of cybersecurity, AI will also be adopted widely in the cybersecurity sector. It should also become obvious that such adoption will lead to AI hacking attacks based on AI properties.

This paper went over the regulatory issues and policy implications for AI policy in cybersecurity by identifying the attracting factors of AI hacker from security related AI properties and various real world cases. It was also pointed out that we need to establish transparency by developing AI regulations from a product safety perspective of requiring strong authentication and record keeping. Also, to undermine the cost-effectiveness of the attack lifecycle and retain minimum external control, we need to implement the non-negligence liability principle and clarify the liability issues. Lastly, this paper proposed the establishment and operation of AI Red Team to identify vulnerabilities through creative penetration testing as well as expand the vulnerability disclosure policy to build the basis of cybersecurity practice.

We need to look into practical cases of AI security and technical trends of cyberattacks to further analyze the attracting factors of AI hacker and its behaviors to identify and deal with relevant issues associated with AI technology.

References

Benjamins, R., Barbado, A., and Sierra D. 2019. Responsible AI by Design in Practice. In *Proceedings of the Human-Centered AI: Trustworthiness of AI Models & Data (HAI) track at AAAI Fall Symposium*.

Brundage, M. et al. 2018. *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. Future of Humanity Institute, University of Oxford, Centre for the Study of Existential Risk, University of Cambridge, Center for a New American Security, Electronic Frontier Foundation, OpenAI: 16-18.

Chung, K. et al., 2021. Machine Learning in the Hands of a Malicious Adversary: A Near Future If Not Reality. In *Game Theory and Machine Learning for Cyber Security*. Wiley-IEEE Press: 291-293.

Ciancaglini, V., Gibson, C., and Sancho, D. 2020. *Malicious Uses and Abuses of Artificial Intelligence*. Trend Micro Research, UNICRI, EC3: 7-9.

Committee on commerce science and transportation. 2014. A "Kill Chain" Analysis of the 2013 Target Data Breach. *Majority Staff Report for chairman Rockefeller*: 7-11.

Darktrace. 2018. The Next Paradigm Shift : AI-Driven Cyber-Attack.: 2-4.

Dempsey, J. X., and Grotto, A. J. 2021. Vulnerability Disclosure and Management for AI/ML Systems: A Working Paper with Policy Recommendation. Stanford Cyber Policy Center: 28.

European Commission. 2021. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence(Artificial Intelligence Act) and Amending Certain Union Legislative Act(COM/2021/206 final). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206 (Accessed: 2022/02/01).

Gerstner, M. E. 1993. Liability Issues with Artificial Intelligence Software. *Santa Clara Law Review* 33(1): 254-258.

Li, Z. et al. 2018. VulDeePecker: A Deep Learning-Based System for Vulnerability Detection. *Network and Distributed Systems Security (NDSS) Symposium* : 3-13.

Maillart, T., Zhao, M., Grossklangs, J., and Chuang, J. 2017. Given Enough Eyeballs, All Bugs are Shallow? Revisiting Eric Raymond with Bug Bounty Programs. *Journal of Cybersecurity* 3(2): 87.

McKinney, D. 2008. New Hurdles for Vulnerability Disclosure. *IEEE Security & Privacy* 6(2): 76.

Mirsky, Y. et al. 2021. The Threat of Offensive AI to Organizations. *ACM Computing Surveys* 1(1): 10.

Sayler, K. M. 2020. Artificial Intelligence and National Security. Congressional Research Service: 34.

Schneier, B. 2021. *The Coming AI Hackers*. Harvard Kennedy School Belfer Center: 41.

Schmit, E. 2021. *Final Report*. National Security Commission on Artificial Intelligence: 641.

Sewak, M., Sahay, S. K., and Rathore, H. 2021. ADVERSARIALuscator: An Adversarial-DRL based Obfuscator and Metamorphic Malware Swarm Generator. *2021 International Joint Conference on Neural Networks (IJCNN)*: 12-17.

Taddeo, M., and Floridi. L. 2019. How AI can be a Force for Good. *Science* 361(6404): 751-752.

Taddeo, M., McCutcheon, T., and Floridi, L. 2019. Trusting Artificial Intelligence in Cybersecurity is a Double-Edged Sword. *Nature Machine Intelligence* 1.

Takaesu, I. and Terada, T., 2016. SAIVS - Spider Artificial Intelligence Vulnerability Scanner. *Black Hat Asia 2016 Arsenal*: 1-2.

Yuehong. 2021. Discussion on the realization technology of virtual hacker robot. *2021 International Conference on Artificial Intelligence and Electromechanical Automation (AIEA)*: 76-77

Wang, Y., Wu, Z., Wei, Q., and Wang, Q. 2019. NeuFuzz: Efficient Fuzzing With Deep Neural Network. *IEEE Access* 7: 36342-36344.

Wang, X., Hu, C., Ma, R., Li, B., and Wang, X. 2020. LAFuzz: Neural Network for Efficient Fuzzing. *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)* : 603-609.