


Editorial

# Foreword to the Special Issue on Advances in Secure AI: Technology and Applications

Sangkyun Lee 

School of Cybersecurity, Korea University, Seoul 02841, Korea; sangkyun@korea.ac.kr

I am pleased to introduce the Special Issue on “Advances in Secure AI: Technology and Applications”. Artificial intelligence (AI) has grown as a key technology that enables numerous applications that facilitate our daily lives. However, this growth has set forth concerns about the safety, security, and reliability of the technology to be used for applications where the malfunctioning of AI can damage human beings or critical infrastructure. Since the findings of adversarial examples in deep neural networks [1], researchers have found many more ways to find adversarial examples (for example, FGSM [2], DeepFool [3], SBA [4], CW [5], and UAP [6], to name a few) and new types of vulnerabilities of deep neural networks such as model stealing [7,8] and data poisoning/backdoor attacks [9,10]. Additionally, it has been discovered that adversarial examples can work in physical environments [11]; for instance, making computer vision systems in autonomous vehicles recognize traffic signs incorrectly [12]. Secure AI is now studied in various fields of computer science beyond the AI community. For example, model stealing can also be done by side-channel attacks [13] and, therefore, would require considerations in secure computing based on multi-party computation [14] and hardware enclaves (e.g., Intel’s SGX and ARM’s TrustZone). Additionally, when distributed learning is preferred due to the cost and security issues of maintaining big data centers, the security of AI may have to be considered in the presence of federated learning [15] or homomorphic encryption [16], probably hardening data privacy via differential privacy [17]. Despite the recent discoveries, it is still not very well understood why certain weaknesses of AI models exist and how to strengthen them against specific exploits entirely. Moreover, as more AI models become available, testing them for vulnerabilities will be essential for trusting and using AI in mission-critical systems, where techniques like XAI (explainable AI) [18] could be applied to inspect logical or statistical issues in AI models.

**Funding:** This research has been supported by a Korea University Grant.

**Conflicts of Interest:** The author declares no conflict of interest.



**Citation:** Lee, S. Foreword to the Special Issue on Advances in Secure AI: Technology and Applications. *Appl. Sci.* **2022**, *12*, 10015. <https://doi.org/10.3390/app121910015>

Received: 28 September 2022

Accepted: 30 September 2022

Published: 5 October 2022

**Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## References

1. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.J.; Fergus, R. Intriguing properties of neural networks. In Proceedings of the 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, 14–16 April 2014.
2. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. In Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
3. Moosavi-Dezfooli, S.M.; Fawzi, A.; Frossard, P. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2574–2582.
4. Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z.B.; Swami, A. Practical Black-Box Attacks against Machine Learning. In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, Abu Dhabi, United Arab Emirates, 2–6 April 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 506–519.
5. Carlini, N.; Wagner, D. Towards evaluating the robustness of neural networks. In Proceedings of the In Proceedings of the 2017 IEEE Symposium on Security and Privacy, San Jose, CA, USA, 22–26 May 2017; pp. 39–57.
6. Moosavi-Dezfooli, S.M.; Fawzi, A.; Fawzi, O.; Frossard, P. Universal Adversarial Perturbations. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 86–94.
7. Tramèr, F.; Zhang, F.; Juels, A.; Reiter, M.K.; Ristenpart, T. Stealing Machine Learning Models via Prediction APIs. In Proceedings of the 25th USENIX Conference on Security Symposium, Austin, TX, USA, 10–12 August 2016; pp. 601–618.
8. Orekondy, T.; Schiele, B.; Fritz, M. Knockoff Nets: Stealing Functionality of Black-Box Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
9. Chen, X.; Liu, C.; Li, B.; Lu, K.; Song, D. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. *arXiv* **2017**, arXiv:1712.05526.
10. Muñoz González, L.; Biggio, B.; Demontis, A.; Paudice, A.; Wongrassamee, V.; Lupu, E.C.; Roli, F. Towards Poisoning of Deep Learning Algorithms with Back-Gradient Optimization. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, Dallas, TX, USA, 3 November 2017; pp. 27–38.
11. Sharif, M.; Bhagavatula, S.; Bauer, L.; Reiter, M.K. Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, 24–28 October 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 1528–1540.
12. Eykholt, K.; Evtimov, I.; Fernandes, E.; Li, B.; Rahmati, A.; Xiao, C.; Prakash, A.; Kohno, T.; Song, D. Robust Physical-World Attacks on Deep Learning Visual Classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
13. Wolf, S.; Hu, H.; Cooley, R.; Borowczak, M. Stealing Machine Learning Parameters via Side Channel Power Attacks. In Proceedings of the 2021 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), Tampa, FL, USA, 7–9 July 2021; pp. 242–247.
14. Ben-Or, M.; Goldwasser, S.; Wigderson, A. Completeness Theorems for Non-Cryptographic Fault-Tolerant Distributed Computation. In Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing, New York, NY, USA, 16–18 May 1988; Association for Computing Machinery: New York, NY, USA, 1988; pp. 1–10.
15. Konečný, J.; McMahan, B.; Ramage, D. Federated Optimization: Distributed Optimization Beyond the Datacenter. *arXiv* **2015**, arXiv:1511.03575.
16. Gentry, C. Fully Homomorphic Encryption Using Ideal Lattices. In Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing, Bethesda, MD, USA, 31 May–2 June 2009; Association for Computing Machinery: New York, NY, USA, 2009; STOC '09, pp. 169–178.
17. Dwork, C. Differential Privacy. In *Proceedings of the Automata, Languages and Programming*; Bugliesi, M., Preneel, B., Sassone, V., Wegener, I., Eds.; Springer: Berlin/Heidelberg, Germany, 2006; pp. 1–12.
18. Gunning, D. *Explainable Artificial Intelligence (XAI), DARPA-BAA-16-53*; Defense Advanced Research Projects Agency: Arlington County, VA, USA, 2016.