KHAZANAH
RESEARCH
INSTITUTE

VIEWS 64/24 | 31 DECEMBER 2024

Cybersecurity Risks of AI

Laventhen Sivashanmugam and Jun-E Tan

Introduction

Artificial Intelligence (AI) as an emergent technology is progressing at breakneck speed. Across the globe, we are starting to see AI systems [1] being trialled in critical infrastructure such as hospitals, financial services and public services, to make these systems more efficient and effective[2].

As the adoption of AI expands, cybersecurity concerns have also increased. These mainly come from the perspectives of how AI can be used to undermine cybersecurity, such as automating or optimising cyberattacks, or conducting fraud and identity theft to obtain unauthorised access. Much less discussed is the aspect of how AI systems can be compromised by malicious actors, affecting the systems' efficacy or using them as entry points to commit data breaches[3].

Views are short opinion pieces by the author(s) to encourage the exchange of ideas on current issues. They may not necessarily represent the official views of KRI. All errors remain the authors' own.

This view was prepared by Laventhen Sivashanmugam, an independent researcher, and Dr Jun-E Tan, a researcher from the Khazanah Research Institute (KRI). The authors are grateful for the valuable comments from Dr Rachel Gong, Gregory Ho Wai Son, Nik Syafiah Anis Nik Sharifulden, Khoo Wei Yang and Salbiah Idris.

Authors' email address:
laventhens@gmail.com and
june.tan@krinstitute.org

Attribution – Please cite the work as follows: Laventhen Sivashanmugam and Jun-E Tan. 2024. Cybersecurity Risks of AI. Kuala Lumpur: Khazanah Research Institute. License: Creative Commons Attribution CC BY 3.0.

Photo by Olemedia on iStock

Information on Khazanah Research Institute publications and digital products can be found at www.KRInstitute.org.

[1] In this article, we follow the OECD definition of an AI system (Russell, Grobelnik, and Perset - 2024): "An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment."
[2] Sakhnini et al. (2020)
[3] Gibian (2022)

In this article, we consider vulnerabilities in AI systems, a topic that is increasingly relevant as our reliance on the systems grows. Broadly speaking, an AI system "learns" from the data we feed into it and makes predictions based on that data. It can learn wrongly, it can be tricked, and it can be reverse engineered. These three issues map onto three AI hacking methods that we will go into:

1) **Data poisoning** affects how an AI model learns, thus reducing its predictive accuracy.
2) **Evasion attacks** attempt to trick an AI model, invalidating the model's predictions.
3) **Model inversion** is when hackers try to reverse-engineer an AI model, to try to steal the model itself, or steal the model's training data.

This article is the third within our AI risk series. Readers who are interested can refer to the first article, "Introducing the AI Risk Series" for the baseline assumptions and understanding of AI[4], and the second on the risks of uneven AI adoption by MSMEs[5].

AI Learning Wrongly: Data Poisoning

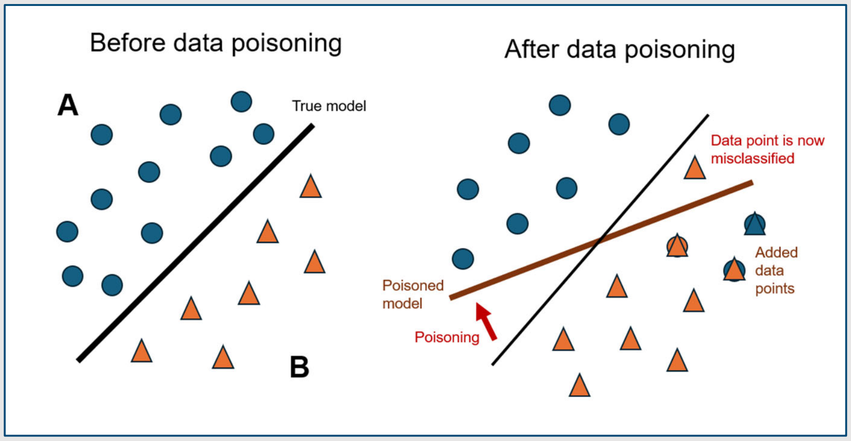
Data poisoning occurs when a hacker feeds tampered data into an AI system, forcing it to "behave the way the attacker wants, as opposed to its creator's intent[6]." As previously mentioned, data poisoning affects the predictive accuracy of an AI model, thus affecting its ability to classify data and make correct decisions[7].

By tampering with training dataset through injecting mislabelled or malicious data, a hacker "teaches" the AI model to behave in a way that benefits the hacker. Box 1 contains a visual representation of data poisoning to help us understand how a model can be "taught" to behave differently. The data points fed into the system affect the model's accuracy, thus rendering it incapable of making correct classifications. For example, Google has previously admitted that its email spam filter has been tricked before. By repeatedly marking malicious spam emails as non-spam with thousands of burner accounts, hackers eventually "taught" the spam filter algorithm to mark spam emails as legitimate[8].

[4] Tan (2024)
[5] Gong (2024)
[6] Gibian (2022)
[7] Chaalan et al. (2024)
[8] Joshi (2022)

Box 1 Diagrammatic Representation of Data Poisoning

Figure 1 Feature space classification depicting the process of data poisoning



Visualisation created by authors, adapted from Miller, Xiang, and Kesidis (2019)

Figure 1 is a modified version of the feature space classification offered by Miller, Xiang, and Kesidis (2019). In the first diagram, the AI model (represented by the black line) can classify the data points (represented by the orange triangles and blue circles) into groups A and B. After data poisoning, the hacker then feeds manipulated data points into the model gradually skews the model's outputs away from the originally accurate classifications. Eventually, the model becomes fully poisoned and starts misclassifying data points into wrong groups.

Another example of real-world implications of data poisoning is provided as follows. Company A uses an AI system to filter out fake business enquiries they receive on their website to save themselves some time. A rival company (Company B) finds out about this and starts spamming Company A's website with fake enquiries and nonsense text but uses the names of real potential customers. The AI system eventually starts to filter out real enquiries from those potential customers in the name of efficiency, and Company A loses business. Applied to a scenario of government agencies receiving many reports or enquiries a day, this also illustrates how public services using AI-based input filters can be hacked.

The strength of an AI model lies in its ability to "learn", but that is also an entry point for hackers to exploit. An AI model reinforces its learning to improve; but errors can become embedded into the model's functions and have far-reaching effects. In the same way a shaky foundation results in a shaky building, data poisoning will affect a model's ability to engage with new data objectively, reinforcing its own inaccurate conclusions. Integrating AI systems into critical

[9] Miller, Xiang, and Kesidis (2019)

infrastructure must therefore be done with a high level of quality control at every stage of development.

AI Being Tricked: Evasion Attacks and Adversarial Examples

The goal of an evasion attack is to mislead an AI model in its classifications, affecting the validity of its predictions and thus the validity of its decisions. To make a distinction between data poisoning and evasion attacks, a data poisoning attack goes after the initial training data in hopes of skewing the AI model itself; an evasion attack is not trying to change the model per se but is instead trying to "trick" it. A hacker can trick an AI model by changing the input in a specific way, so the model misreads the input, and the system takes the wrong course of action[10].

Here is an example to illustrate the process. An image recognition AI system assesses every pixel in an image, assigns markers based on the data it's trained on, and classifies images accordingly. As the statistical analysis an AI model undertakes to classify images is complex, we may not immediately understand what the markers are. Take for example a case presented during an international conference with the Association of Computing Machinery, where a group of researchers trained an image classifier AI model to differentiate between a wolf and a dog. The model succeeded at classifying images correctly, but after analysing its processes, the researchers realised that it was differentiating dogs and wolves based on the presence of snow in the images instead of the characteristics of the animals themselves[11].

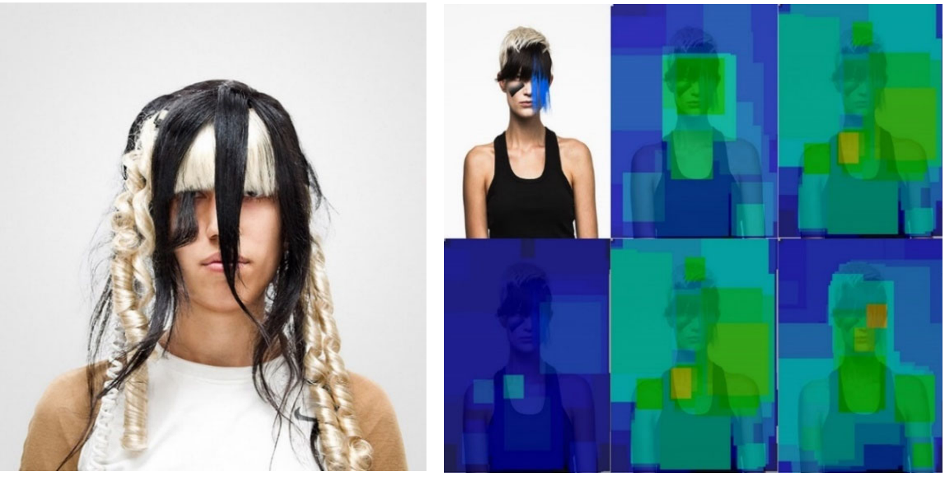
Someone who understands how an AI model classifies objects can exploit them. In 2010, Adam Harvey, a postgraduate student at New York University, discovered unusual ways of tricking computer vision algorithms with hairstyles and makeup. Figure 2 shows a few examples of these looks, which strategically cover up parts of the face that face detection models would immediately recognise; elements such as colour, shapes and facial asymmetry also played a part in his looks. With trial-and-error and an understanding of how computer vision algorithms process light and shape, Harvey was able to make his test subjects immune to facial detection AI models[12].

[10] IBM (2024)

[11] Ribeiro, Singh, and Guestrin (2016)

[12] 'CV Dazzle' (2011)

Figure 2: Makeup and hairstyles that can trick face detection algorithms



Source: Adam Harvey

Evasion attacks can be employed in both physical and digital worlds. An example in the physical world would be attacks on self-driving cars. The autonomous driving systems of these cars can misread traffic signals and thus make the wrong driving choices, potentially resulting in car accidents[13]. In the digital world, evasion attacks can trick AI-powered detection systems, with known cases of insurance fraud in the insurance industry[14] or offensive or problematic content bypassing automated content moderation[15].

AI Being Reverse-Engineered: Model Inversion, Model Extraction and Membership Inference

There are two possible objectives of a model inversion attack: a hacker is either interested in stealing the target AI model itself or stealing the data that they can obtain through the target model. A hacker conducting an attack engages directly with an AI system by feeding it information and seeing its behaviour and figuring out its parameters via that behaviour[16], a bit like feeling the shape of a present over the wrapping paper. If the hacker wants to steal the model itself, this is known as **model extraction**. If the hacker wants to steal or retrieve training data of the model, this is known as **membership inference**.

[13] Kong et al. (2020)
[14] Cattaneo, Kenett, and Luciano (2024)
[15] Tambe et al. (2021)
[16] He, Zhang, and Lee (2019)

Model Extraction: Stealing Models

If a hacker has access to a target model's training data (e.g. it is publicly available or the dataset has been leaked) and can guess a model's parameters from its behaviour , they can try to replicate the model[17]. With a replicated model, the hacker can then accomplish other objectives such as creating and testing adversarial examples for evasion attacks or copy an AI system without permission and compete against the business selling the original product.

A real-life example of the risks of model extraction can be seen in the case of cybersecurity. Modern cybersecurity mechanisms employ AI systems to detect vulnerabilities or threats, such as viruses or firewall breaches[18]. An AI-powered cybersecurity system can detect these threats automatically and quickly, without needing constant monitoring by a cybersecurity expert[19]. However, if a hacker can replicate the system, its threat detection abilities can be circumvented and become obsolete. With the replicated system, the hacker can create an evasion attack tailor-made to trick the target cybersecurity AI system. Once the cybersecurity model has been "extracted", the hacker would know what kinds of code/files the system registers as malware and would formulate the exact code/virus that would slip past its defences.

Membership Inference: Stealing Data

Membership inference refers to the process of inferring characteristics about the training data based on the model's output, which presents security and privacy risks for sensitive data. To put it in broad terms, the AI model creates its internal architecture (such as weights and optimisers) based on the data it receives. By understanding the architecture of the model, a hacker can potentially reverse engineer the data that it was originally trained on.

The health sector is often mentioned as a site for membership inference, given the amount of sensitive patient data held in healthcare facilities or medical databases for research[20]. However, high value and confidential data can range widely from business trade secrets to governmental data for national security, beyond sensitive personal data. There are options to safeguard against membership inference attacks such as employing differential privacy measures, which inserts noise or randomness into datasets, but these techniques can be expensive for large datasets[21].

[17] Erdoğan, Küpçü, and Çiçek (2022)
[18] Camacho (2024)
[19] Ansari et al. (2022)
[20] Davenport and Kalakota (2019)
[21] Gibian (2022)

Conclusion

As AI adoption expands, cybersecurity concerns of AI systems should receive commensurate attention. Data poisoning and evasion attacks are well documented within the cybersecurity world, but threats like membership inference are still quite new and mainly exist as proofs of concept in academic papers[22]. Even so, the possibilities of AI systems being hacked need to be taken seriously, given the speed and scale of harms that can be perpetuated using automated decision-making systems.

With these security vulnerabilities in mind, what do potential solutions look like?

- Data poisoning is ultimately a data security issue and therefore countermeasures such as data quality control, routine auditing and continuous monitoring can be very effective[23].
- Evasion attacks can be defended against by training our AI models to recognise potential evasion attacks[24], or by building AI systems with multiple AI models as components, so that hackers have a harder time slipping through[25].
- Model inversion is a threat because it allows hackers to glean sensitive information by making inferences about the output of the target model; differential privacy prevents this from happening by obfuscating the inferences themselves, making it very difficult for hackers to “steal” your private information[26].

Ultimately, the solutions boil down to being intentional with AI implementation, conscientious with data, and sparing no effort during the AI audit process. Security is especially pertinent in the case of AI being implemented in national critical infrastructure and systems involving public interest. State actors as well as businesses looking to maximise efficiency and effectiveness using AI must not only implement data governance policies that ensure security and privacy, but also invest in cybersecurity measures that cover AI system vulnerabilities.

[22] DeAlcala et al. (2024)
[23] Ferrag et al. (2024)
[24] Zhao, Alwidian, and Mahmoud (2022)
[25] Ahmed, Lin, and Srivastava (2022)
[26] Ye et al. (2022)

References

Ahmed, Usman, Jerry Chun-Wei Lin, and Gautam Srivastava. 2022. ‘Mitigating Adversarial Evasion Attacks of Ransomware Using Ensemble Learning’. *Computers and Electrical Engineering* 100 (May):107903. https://doi.org/10.1016/j.compeleceng.2022.107903.

Ansari, Meraj Farheen, Bibhu Dash, Pawankumar Sharma, and Nikhitha Yathiraju. 2022. ‘The Impact and Limitations of Artificial Intelligence in Cybersecurity: A Literature Review’. SSRN Scholarly Paper. Rochester, NY. https://papers.ssrn.com/abstract=4323317.

Camacho, Nicolas Guzman. 2024. ‘The Role of AI in Cybersecurity: Addressing Threats in the Digital Age’. *Journal of Artificial Intelligence General Science (JAIGS) ISSN: 3006-4023* 3 (1):143–54.

Cattaneo, Matteo, Ron S. Kenett, and Elisa Luciano. 2024. ‘Adversarial AI in Insurance: An Overview’. *European Actuarial Journal* 14 (1). Springer:297–306.

Chaalan, Tarek, Shaoning Pang, Joarder Kamruzzaman, Iqbal Gondal, and Xuyun Zhang. 2024. ‘The Path to Defence: A Roadmap to Characterising Data Poisoning Attacks on Victim Models’. *ACM Computing Surveys* 56 (7):175:1-175:39. https://doi.org/10.1145/3627536.

‘CV Dazzle’. 2011. 1 September 2011. https://adam.harvey.studio/cvdazzle/.

Davenport, Thomas, and Ravi Kalakota. 2019. ‘The Potential for Artificial Intelligence in Healthcare’. *Future Healthcare Journal* 6 (2):94–98. https://doi.org/10.7861/futurehosp.6-2-94.

DeAlcala, Daniel, Aythami Morales, Julian Fierrez, Gonzalo Mancera, Ruben Tolosana, and Javier Ortega-Garcia. 2024. ‘Is My Data in Your AI Model? Membership Inference Test with Application to Face Images’. arXiv. https://doi.org/10.48550/arXiv.2402.09225.

Erdoğan, Ege, Alptekin Küpçü, and A. Ercüment Çiçek. 2022. ‘UnSplit: Data-Oblivious Model Inversion, Model Stealing, and Label Inference Attacks against Split Learning’. In *Proceedings of the 21st Workshop on Privacy in the Electronic Society*, 115–24. Los Angeles CA USA: ACM. https://doi.org/10.1145/3559613.3563201.

Ferrag, Mohamed Amine, Fatima Alwahedi, Ammar Battah, Bilel Cherif, Abdechakour Mechri, and Norbert Tihanyi. 2024. ‘Generative AI and Large Language Models for Cyber Security: All Insights You Need’. arXiv. https://doi.org/10.48550/arXiv.2405.12750.

Gibian, Davey. 2022. *Hacking Artificial Intelligence: A Leader’s Guide from Deepfakes to Breaking Deep Learning*. https://rowman.com/ISBN/9781538155080/Hacking-Artificial-Intelligence-A-Leaders-Guide-from-Deepfakes-to-Breaking-Deep-Learning.

Gong, Rachel. 2024. ‘Not Leaving MSMEs Behind In The AI Race’. Kuala Lumpur: Khazanah Research Institute. https://www.krinstitute.org/Views-@-Not_Leaving_MSMEs_Behind_In_The_AI_Race__.aspx.

He, Zecheng, Tianwei Zhang, and Ruby B. Lee. 2019. ‘Model Inversion Attacks against Collaborative Inference’. In *Proceedings of the 35th Annual Computer Security Applications Conference*, 148–62. San Juan Puerto Rico USA: ACM. https://doi.org/10.1145/3359789.3359824.

IBM. 2024. 'Evasion Attack Risk for AI'. 3 October 2024. https://www.ibm.com/docs/en/watsonx/saas?topic=atlas-evasion-attack.

Joshi, Naveen. 2022. 'Countering The Underrated Threat Of Data Poisoning Facing Your Organization'. Forbes. 21 March 2022. https://www.forbes.com/sites/naveenjoshi/2022/03/17/countering-the-underrated-threat-of-data-poisoning-facing-your-organization/.

Kong, Zelun, Junfeng Guo, Ang Li, and Cong Liu. 2020. 'PhysGAN: Generating Physical-World-Resilient Adversarial Examples for Autonomous Driving'. In , 14254–63. https://openaccess.thecvf.com/content_CVPR_2020/html/Kong_PhysGAN_Generating_Physical-World-Resilient_Adversarial_Examples_for_Autonomous_Driving_CVPR_2020_paper.html.

Miller, David J., Zhen Xiang, and George Kesidis. 2019. 'Adversarial Learning in Statistical Classification: A Comprehensive Review of Defenses Against Attacks'. arXiv. https://doi.org/10.48550/arXiv.1904.06292.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. '"Why Should I Trust You?": Explaining the Predictions of Any Classifier'. arXiv. https://doi.org/10.48550/arXiv.1602.04938.

Russell, Stuart, Marko Grobelnik, and Karine Perset. 2024. 'What Is AI? Can You Make a Clear Distinction between AI and Non-AI Systems? - OECD.AI'. 6 March 2024. https://oecd.ai/en/wonk/definition.

Sakhnini, Jacob, Hadis Karimipour, Ali Dehghantanha, and Reza M. Parizi. 2020. 'AI and Security of Critical Infrastructure'. In *Handbook of Big Data Privacy*, edited by Kim-Kwang Raymond Choo and Ali Dehghantanha, 7–36. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-38557-6_2.

Tambe, U.S., N.R. Kakad, S.J. Suryawanshi, and S.S. Bhamre. 2021. 'Content Filtering of Social Media Sites Using Machine Learning Techniques'. In . https://doi.org/10.3233/APC210226.

Tan, Jun-E. 2024. 'Introducing the AI Risk Series'. Kuala Lumpur: Khazanah Research Institute. https://www.krinstitute.org/Views-@-Introducing_the_AI_Risk_Series.aspx.

Ye, Dayong, Sheng Shen, Tianqing Zhu, Bo Liu, and Wanlei Zhou. 2022. 'One Parameter Defense -- Defending against Data Inference Attacks via Differential Privacy'. arXiv. https://doi.org/10.48550/arXiv.2203.06580.

Zhao, Weimin, Sanaa Alwidian, and Qusay H. Mahmoud. 2022. 'Adversarial Training Methods for Deep Learning: A Systematic Review'. *Algorithms* 15 (8). Multidisciplinary Digital Publishing Institute:283. https://doi.org/10.3390/a15080283.