The Prevention of Model Poisoning Using Artificial Immune Networks.

Model Poisoning

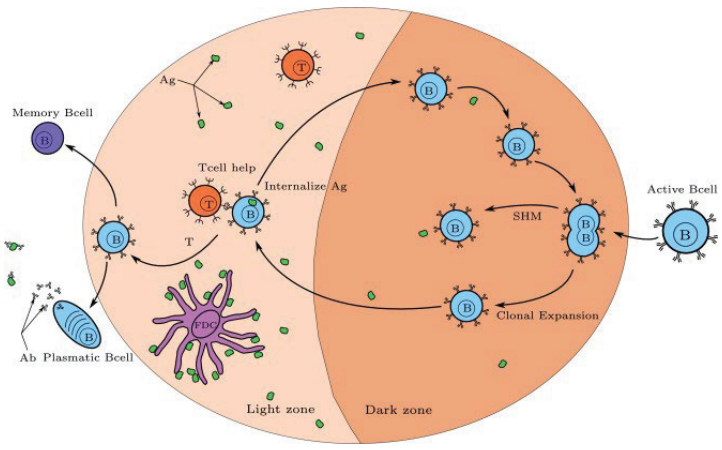
Ashir H Mohabir
217045508
ashir.mohabir@gmail.com
217045508@student.uj.ac.za

UNIVERSITY OF JOHANNESBURG

Why Poison AI Models?

Model Poisoning refers to an attack malicious actors make on Artificial Intelligent systems. This attack targets the data that models use for training and creates back doors and cracks in the AI system. Model poisoning is achieved by sending through slightly tampered data that will register as benign. When the model used in an Intrusion Detection System is poisoned, bad actors can send harmful data through the network, with the detection system incorrectly deeming it as safe. This ultimately allows bad actors to successfully deal damage to the network, access sensitive data, and have constant access to the network.

Can Artificial Immune Systems Serve as the Antidote?



(https://www.sciencedirect.com/topics/immunology-and-microbiology/artificial-immune-system)

Artificial Immune Systems (AIS) are bio-inspired computational models that use ideas and concepts from immune network theory.

Purpose

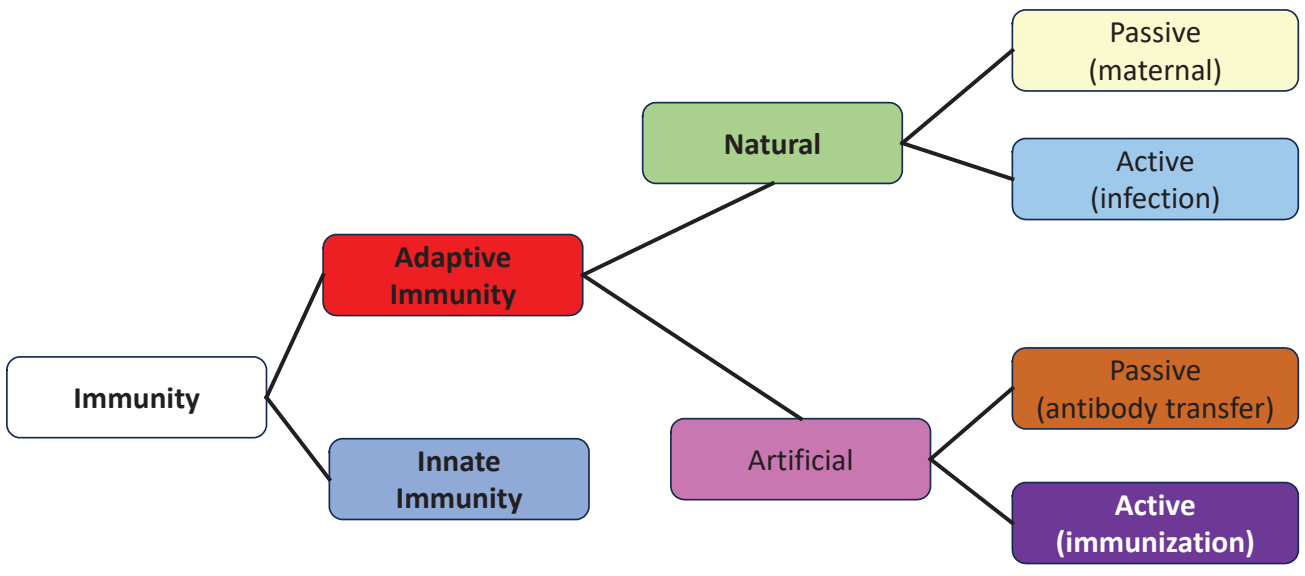
New NISs have started to use Artificial Intelligence (AI) to lessen the burden on network administrators, who would have to constantly create a pre-signature list for the network intrusion systems whenever a new threat was detected. AI makes this more manageable as the AI will detect when new threats are found and update itself to deal with those threats. However, bad actors can target the AI itself. Model poisoning is achieved by sending through slightly tampered data that will register as benign. This concept is known as model poisoning, and it can be dangerous to organisations that use AI to help with their network intrusion systems.

Relevance

The use of machine learning models has become ubiquitous, their predictions are used to make decisions about healthcare, security, investments, and other critical decisions. Given this pervasiveness, it is not surprising that bad actors are incentivised to manipulate machine learning models and find ways to create vulnerabilities in them. One way to manipulating the model is through poisoning, something that has been discussed in previous chapters, but this is when bad actors tamper with the training data that is used by the model to learn. When a model is poisoned, bad actors can manipulate the model, lowering the accuracy rate and increasing the number of false positives going through the system. One way to stop this is by using the concept of immunology. Immunology is a concept that tries to replicate biological aspects in machines and models, in the case of immunology, it is trying to replicate the immune system in machine learning models. This immune system is known as an Artificial Immune System (AIS), and it will act as a protection mechanism for the machine learning model.

Outcome

The creation of an Artificial Immune Network that will then act as a protection mechanism for the machine learning model. The AIS will be protecting the model from model poisoning.



References
1. Rashid N. Iqbal J. Mahmood F. Abid A. Khan US. Tiwana MI.: Artificial Immune Systems-Negative Selection Classification Algorithms (NSCA) for Four Class Electroencephalogram (EEG) signals. (Neuroergonomics: The Brain at Work in Everyday Settings. 2018). (2018).
2. Jim LE. Islam N. Gregory MA.: Enhanced MANET security using artificial immune system-based danger theory to detect selfish nodes. (Computers & Security, 113, 2022, 102538). (2022).
3. Yoshida K. Fujino T.: Disabling Backdoor and Identifying Poison Data by using Knowledge Distillation in Backdoor Attacks on Deep Neural Networks. (AISec'20: Proceedings of the 13th ACM Workshop on Artificial Intelligence and Security, 117-127). (2020).
4. Baracaldo N. Chen B. Ludwig H. Safavi JA.: Mitigating Poisoning Attacks on Machine Learning Models: A Data Provenance Based Approach. (AISec'17: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, 103-110). (2017).
5. Haui M. Cai R. Yao L. Zhang A.: Malicious Attacks against Deep Reinforcement Learning Interpretations. (KDD'20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 472-482) (2020).