Swinburne University of Technology

Faculty of Science, Engineering and Technology

ASSIGNMENT AND PROJECT COVER SHEET

Unit Code: COS30015 Unit Title: IT Security
Assignment number and title: Assignment 1 Research Project Due date: 8th Sept 2022
Lab group: Thursday 8:30AM Tutor: Jamie Ooi Lecturer: Lin lin

Family name: Rezwan Identity no: 103172423

Other names:

To be completed if this is an INDIVIDUAL ASSIGNMENT
I declare that this assignment is my individual work. I have not worked collaboratively, nor have I copied from any other student's work or from any other source except where due acknowledgment is made explicitly in the text, nor has any part been written for me by another person.

Signature: S M Ragib Rezwan

To be completed if this is a GROUP ASSIGNMENT
We declare that this is a group assignment and that no part of this submission has been copied from any other student's work or from any other source except where due acknowledgment is made explicitly in the text, nor has any part been written for us by another person.

ID Number	Name	Signature

Marker's comments:

Total Mark:

Extension certification:

This assignment has been given an extension and is now due on

Signature of Convener: Date: / 2022

Unit Code: COS30015

Unit name: IT Security

Title of the assignment: Research Report On AI

Topic: Emerging Attacks on AI

Author: S M Ragib Rezwan (103172423)

Submission Date: 8th Sept 2022 (at 06:30)

Due Date: 8th Sept 2022 (at 23:59)

ABSTRACT

Artificial Intelligence (AI) is a way to simulate human intelligence in a machine in order to help it execute complex tasks, in an effective manner, via the help of learning, reasoning, etc**[1]**. Thus, as the current technology advances, it's not surprising to see AI being integrated in various places, like in Voice Assistant, Self-driving cars, Automated warehouse management systems, etc.**[2]** in order to upgrade the systems. However, as their use increases over time, so does the variety of attacks made against it, in order to disrupt their services. So, here, I am going to first discuss about the papers and articles written on this matter over the years. Then, I will compare over the various attacks and their solutions proposed, alongside my own thoughts on them, in order to better prepare for the future where AI will be integrated in all parts of the human life.

Key-finding: The fact that most attacks can be considered as a form of input attack being performed at different stages of the system (i.e. middle of development, post-development, etc.) to mislead, corrupt, or overload the AI, and thus disrupt the service (no matter what type) provided by the system (like voice, picture recognition, etc.).

[Note: *In this paper, I have used Machine learning (ML) interchangeably with AI. That's because ML is basically a subset of AI which solves specific tasks by learning from the data provided to it **[3]** and so, the attacks made against it will also have the same impact on AI.*]

INTRODUCTION

From its first official use in the research world via "Turing Test" in 1952, to determine "whether computers can replicate human thinking"**[4]**, to current time, where Artificial Intelligence (AI) is being used to perform even daily functions like personalized shopping, AI has advanced by an enormous amount. So much so that no matter where we look today, we can still find its presence, like in fraud prevention method in credit card, in GPS navigation system, etc.

But, being used widespread has also led to it being integrated into systems used to protect and process various crucial information (like hospital records, transaction details, malware detection, etc.), which in turn made it into a big target for various forms of attacks. After all, these systems are heavily reliant on AI's capabilities in order to perform their tasks and so if any attack prevented the AI from function properly, even if by a small amount, the consequences would be disastrous. We can clearly understand this if we consider self-driving cars where they need to recognize nearby objects and calculate the distance between the car and the objects in real time and even the slightest variation in calculation (especially in high speeds) could result in serious accidents.

Moreover, AI is also currently being used in many cyber defense systems worldwide (like in Botnet detection, antimalware, etc.)**[5]** and the failure in any of them (even for a single case) would have dire consequences for all the organizations utilizing them.

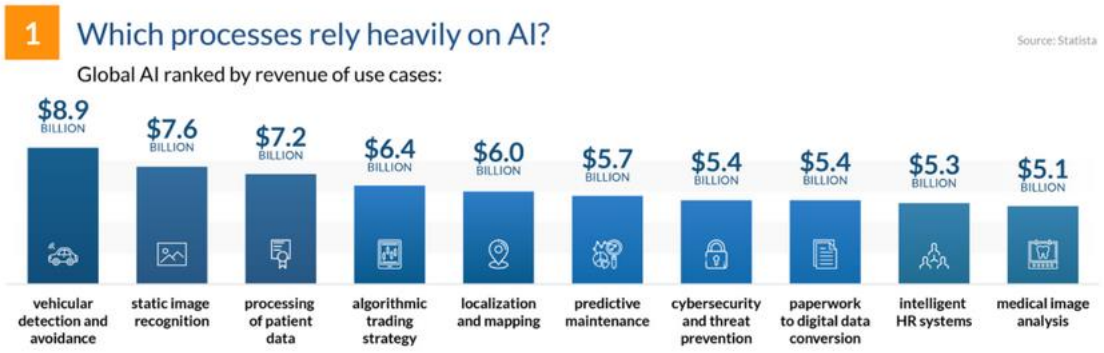

Figure: Processes that are heavily reliant on AI**[6]**

But considering the fact that AI has become integrated in various types of systems, are the attacks made on the AI also fully different depending on which system it is attached to? Or is it more like a "quick change" screw driver (one where you can change the screw driver bits attached) where the method of attack is the same and only the intrusion point changes depending on the system that is being attacked?

From the information I have gathered through the various papers and articles, I have surprisingly found it to be mostly of the second type where even though the attacks are highly specialized (like ones attacking AI used for voice assistants not being able to attack one used for fingerprint detection), that is only the case for their "intrusion point" into the system and the things they do afterwards ends up being exactly same. Thus, I was able to group all of those attacks into a few common types of methods, depending on how they had harmed the AI after gaining entry into the system, which can be seen in the Literature Review and Discussion sections.

Since there had been quite a few Literature topics relevant to "Emerging attacks on AI", I have divided the paper into the following parts to make it easier to understand everything. In the Overview/ background section, I have given a brief overview on how I had selected my information sources and the rationale behind it. In Literature review, I have noted down some of the challenges I had faced in gathering the information, alongside noting the attack types and their existing solutions. In the Discussion section, I have summarized and compared the attack types and their existing solutions and used my own understanding to both rank them in terms of their impact, and also created a manual of general solutions relevant in mitigating all the attacks. Lastly in the Conclusion section, I gave a brief summary about my findings and used my own understandings to note down the weak points in my literature review and ways to improve it for future considerations about the topic.

OVERVIEW/BACKGROUND

Research on AI has been around since 1952 **[4],** and its use in technology since 1978 via SCARA**[7]**, leading to a huge amount of information regarding AI. But unfortunately, most of these focus on how AI can be integrated into either benign or malicious systems (like in developing software for diagnosing infectious diseases **[8]** or even creating Deep-fakes**[9]**), in order to increase their effectiveness. Thus, comparatively, fewer papers focus on the actual attack being made against

the AI systems and the ways to defend against them. So, in order to gather all information regarding this matter, I increased my research radius to include not only peer-reviewed papers from the Swinburne library and Google scholar, but also different articles and blogs released by trustworthy news agencies and top cyber security and technology companies.

But, after that, whilst scanning though all the information I had gathered, I noticed an unusual fact. Although the authors speaking about the attacks were from different region, looking at attacks on different AI systems and at different point in time, most of the attacks they had described were quite similar. Furthermore, some of the attacks were completely same, except noted as different attacks as the writers had given them different names.

Thus, I started to filter and merge the information I had obtained, removing duplicate information, differentiating them via the type of attack (instead of the system they had been attacking on) before finally noting all of them down.

LITERATURE REVIEW

Although there are several variations of attacks on AI, each with their own unique features and attributes, they can be commonly classified under the following types of attacks:

1) **Adversarial Input Type Attack:**

   These types of attacks have been discussed in-depth by both Marcus **[10]** and Sara and her team **[11],** and also briefly touched upon by Alchekov and et al. **[12],** Tom **[5],** Shraddha **[13]** and Elie **[14].** Basically, these attacks focus on altering the input that is being passed to AI system in order to ensure that the outcome serves the attacker's goal. This can be divided into the following sub-types:

   a) Perceivable attacks: Attacks made using inputs that can be observed by people. For instance, this can be simply performed by using a teddy bear where marker can be used in a certain pattern in certain areas to fool the AI picture recognition system into believing that it is not a teddy bear, but something else (like maybe a pillow or a real bear).

   b) Imperceivable attacks: Attacks made using inputs which can't be observed by people. For instance, the Psychoacoustic Hiding technique**[12]** is a way to attack AI voice systems by hiding the commands that only the machine can hear inside audio files.

   c) Physical attacks: Input Attacks made by modifying physical entities. For instance, putting stickers on a turtle to make the AI picture recognition system believe it is a gun and not a turtle.

   d) Digital attacks: Input Attacks made by modifying digital. For instance, this can be done by modifying a picture by using editing device to make the AI picture detection

system believe it is something else (like the picture can be that of a dog, but by putting glasses on it the system may consider it to be that of a person)

Furthermore, the attacks can also be categorized into two types depending on amount of information known by the attacker:

i. Black box: attacks made on the system with little knowledge on how it works
ii. White box: attacks made on the system with full knowledge on how it works

[**Note**: *Although the input attack can be classified into these 6 subtypes, these subtypes are not mutually exclusive. Thus, in most input attacks, a mixture of subtypes can be noticed. For instance, Fast Gradient Signed Method (FGSM)**[11][15][17]** is an attack made by adding small distortions(unseen by human eye) onto a digital images to attack or confuse any image recognition system (like ones used for medical imaging**[11]**). This is an example of imperceivable, digital and white box subtypes of the input type attack*]

These can be resolved in the following ways:

- Using Chow's method of Denosing and verifying Cross-Layer ensemble**[16]** to enable the system automatically detect adversarial attack and repair itself by removing the "noise" or modification attached to the input,
- Training AI against all the attacks from Adversarial Inputs (like training it using robust DataOps (data Operations) or MLOps (Machine Learning Operations) solutions)**[22],**
- Monitoring and rectifying the inputs before passing them to AI system,
- Ensuring that only authorized people know how the entire system works, etc.

2) Poisoning Type Attack:

These types of attacks have been discussed indepth by both Marcus **[10]** and Shraddha **[13],** and also briefly Tom **[5]** and Elie **[14].** Basically these are similar to input attacks, except these are performed during the development stage of the AI system whilst input attacks are only performed after AI system has been fully developed. This can further lead to causing a Distributed denial of service (DDOS) in the future or make way for Trojan attacks!

The initial attack can be divided into the following subtypes:

a) Dataset Poisoning: This can be done by providing an incorrect or mislabeled data to the dataset of the AI system. Since AI learns by recognizing patterns, this will end up misguiding it and thus end up developing a poisoned AI system. The simplest way to do this is by using a label flipping attack **[17]** where the attacker simply "flip" the labels of some of the dataset inputted into training the system.

b) Algorithm Poisoning: This can be done by utilizing a weakness in the Algorithm system used by AI system to learn. This had been noticed in the Federated Learning

algorithm system **[10] [25]** where instead of centrally collecting all data and then training the system, a different method had been used. There, small algorithm models will be trained on the user's device using the user's provided data, before combining them all into the final AI system model. This allowed the user to be able to poison the algorithm model and thus end up poisoning the AI system.

c) Model poisoning: This can be done by either modifying the entire model file of the AI system or just replacing the actual model with a poisoned one during the development stage. For instance, the malicious actor can use an AI model builder **[18]** or other tools to create a poisoned AI model and then replace the original AI model file of the system.

These can be resolved in several ways:

- Training Data filtering techniques like input manipulation detection and gradient shaping **[19],**
- Using Robust Learning like model robustifying and model verification **[19],**
- Using auxiliary tools like GAN (Generative adversarial network) and Robust Statistics **[19]**,
- Ensuring only authorized people have access to the system,
- Constantly monitoring the inputted data and using proper prevention methods, etc.

3) **Model Attacks**: These types of attacks have been discussed by Vijaysinh **[20]** and Qianqian and his team **[21],** and also briefly Henry **[22]** and Tom **[5].** Basically these types focus on attacking the Model used by the AI system after it has been developed. These can be further divided into two subtypes:
   a) Transfer learning or model stealing: This is similar to model poisoning attack except that the attack is performed after the model is developed and here the model is duplicated instead of being modified. This is used to perform data phishing privacy attacks by reverse engineering the dataset in the model and can be done by using different tools. Like for AI system using image classification, Inception, ResNet, etc. **[20]** can be used for "transferring the learning" from the model.

[Note: *Transfer learning is considered a legitimate method used to transfer the learning of a model in order to avoid training a new model from scratch. But since it is a form of "model stealing" (especially in case of malicious actors), I have noted it here as part of an AI attack. Moreover, it can be used to test the effectiveness of attacks on the real model by testing it on the duplicate model*]

   b) AI Model extraction Attack (using SCA (side channel attack) **[21]** on IOT systems): This has been noticed in IOT systems empowered by AI. In those systems, SCA had been used to extract sensitive information of the system (like structure of model, hyperparameters, etc.) via using a security exploit on either the system or the chip hosting the system.

These can be resolved in the following ways:

- Using Deceptive perturbations **[23]** to apply noise to the system and degrade or slow down model stealing attack,
- Using Fuzzy analysis **[21]** and other detection methods to detect Model extraction attacks,
- Ensuring only authorized people have access to the model files of the AI system,
- Using proper protocol and tools to protect access to the core components of the AI system (i.e. using specialized firewalls),
- Using properly fine-tuned model and object functions,
- Using encrypted model and accepting only encrypted input, etc.

4) **Distributed Denial of Service (DDOS) Attacks**: These types of attacks have been discussed by Henry **[22]** and Tom **[5].** Although it is similar to input type attacks, it is not completely same. In input type attacks, the AI system is fed data that it is misleading but understandable for the system. But here, complex data is being sent to it continuously to ensure that the model is unusable for its intended purpose for a certain period of time.

[Note: *Although there hasn't been much documented case of DDOS attacks focusing primarily onto AI systems, there have been several DDOS attacks on systems that had utilized AI. These include Google's system, Amazon's system, etc.* ***[27]****]*

These can be resolved in the following ways:

- Using proper security compliance policies,
- Monitoring and managing the system,
- Setting up proper, upto-date firewalls and using a resilient model architecture,
- Using detection and prevention mechanisms by using specialized AI techniques (like Naïve Bayes or Random forest tree) **[24]** alongside the default AI in the system**,** etc.

[Note: *Here I am* n*ot speaking about DOS attacks as DDOS is just a Distributed DOS attack. Also, I am not considering it as subtype of DDOS attack as DOS: a) is just a weaker version of DDOS, and b) doesn't have as much impact on AI system as DDOS.*]

[Note: *Some literature articles also consider another type of attack termed "Online adversarial attack"* ***[22]*** *where attackers manipulate a model's learning by feeding it false data online. But I have decided not to consider this as a separate attack as it is exactly same as the sub type "dataset poisoning" from "poisoning type attacks" and the poisoning attack types can be both online and offline attacks*]

Discussion

First of all, it's good to keep in mind that most of the literature sources speaking on the topic are not papers but instead are articles or blogs and among those papers, very few are peer-reviewed. Thus the quality and accuracy of the information obtained cannot be ascertained.

Moreover, even though different papers, articles and blogs have been found speaking about Emerging Attacks on AI, their thoughts weren't exactly same. For instance, almost all the literatures accepted input attacks, poisoning attacks as two forms of attacks, but few considered DDOS as also another form of attack on it. Furthermore, some articles even ended up contradicting one another as their focus had been on different aspects (for instance, in Marcus' paper**[10]** Federated learning exploit is considered at algorithm level poisoning attack, while in Lingjuan's book **[25]**, it is considered to be at model level)!

Even though it may feel surprising to see people researching on the similar topics but coming to diverse conclusions, the rationale behind it can actually be understood if we look at the figure**[26]** below. There we can see that by focusing their observation on a specific, narrow area instead of looking at the entire broad object, each researcher have come to different conclusions about the exactly same object!

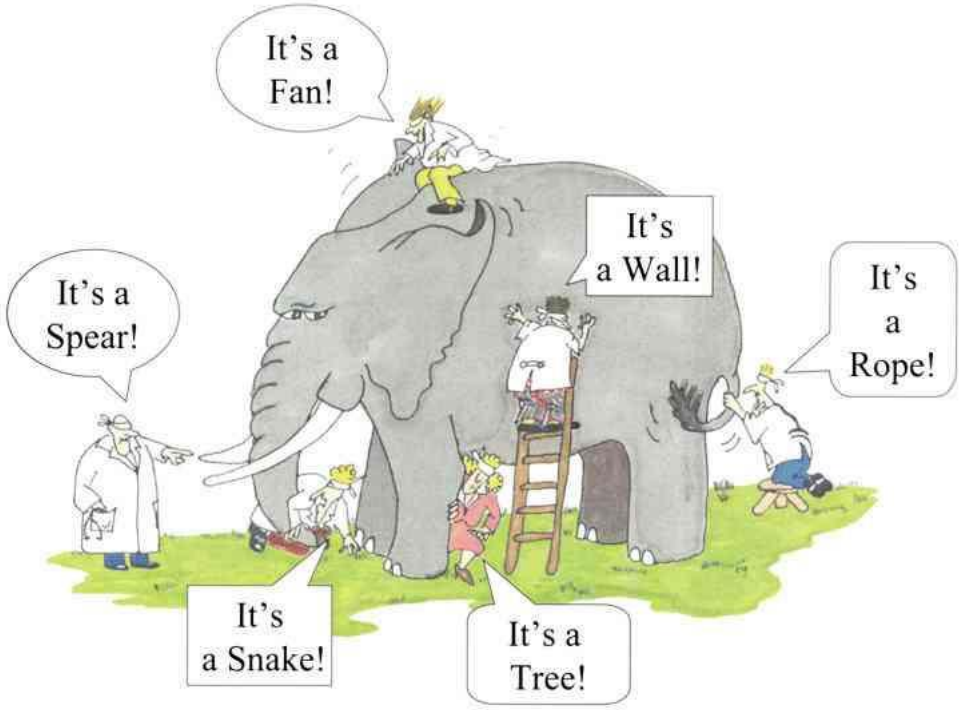

Figure: blind-folded researchers coming to different conclusions for the same object **[26]**

So, in order to get a better understanding of the information obtained and hence both rank the attacks against AI and also find common solutions to them, I have decided to compare the attacks using the following table:

Attack type	**Attack subtype**	**Some examples of the attack**	**Mitigation methods**
Input Type	Perceivable, Impercieveable, Physical and Digitial	FGSM (fast gradient sign method), PDG (projected	Using Chow's method **[16]** Training AI using robust DataOps or MLOps**[22]**

		gradient descent), Cand W(Carlini and Wagner) method **[11]**, street signs modification**[10]**, etc.	Monitoring and rectifying the inputs Restricting system access to authorized people only
Poison Type	Dataset poisoning, Algorithm poisoning, Model poisoning	Label flipping, clean label data poisoning, backdoor attack,etc.	Training input manipulation detection and gradient shaping **[19]** Using model robustifying and verification **[19]** Using GAN and Robust Statistics **[19]** Restricting system access to authorized people only Monitoring inputs and using prevention methods
Model Attack	Transfer learning/ Model stealing, Model extraction attack	Model Extraction attack via SCA, ResNet, etc.	Using Deceptive perturbations **[23]** Using Fuzzy analysis **[21]** Restricting system access to authorized people only Using specialized firewalls Using properly fine-tuned model and object functions Using encrypted model and accepting only encrypted input
DDOS Attack	No subtype	Although examples of directly DDOSing an AI has not been found, there have been several examples of systems(that use AI in them) being DDOSed, like DDOS attack on Amazon, Google, etc. **[27]**	Using proper security compliance policies Monitoring and managing the system Setting up proper, upto-date firewalls Using a resilient model architecture Using Naïve Bayes or Random forest tree**[24]**

From here we can see that input and poisoning types attacks had the higher number of examples on directly attacking AI systems, compared to model attacks. Moreover, although DDOS attacks were far more common, they didn't directly focus on attacking the AI in the system and instead attacked the entire system. Thus, in terms of ranking the attacks on occurrence, input and poisoning type are at the top, followed by model attack and ending with DDOS types.

After this I decided to remove some of the general or common mitigation techniques from the solutions in order to make:

a) List of steps/manual that should be followed, in general, to protect from all forms of attack:
    i. Monitoring and rectifying not only the inputs entered into system but also the entire system,
    ii. Restricting system access to authorized people only,
    iii. Using specialized firewalls and other security compliance policies,
    iv. Encrypting both inputs passed and the system as whole.

b) Comparing the mitigation techniques to see which ones use a higher number specialized/complex tools to create another ranking:

Attack type	**No. of specialized tools used**
Input type	3
Poison type	6
Model attack	2
DDOS	2

Here, we can see Poison type attacks leading with 6 specialized tools, followed by input types attacks with 3 specialized tools and ending with model attack and DDOS, both of which have 2 specialized tools to mitigate the attacks.

Thus, I decided to merge the mitigation ranking with results of the occurrence ranking, keeping the ones with both higher relevant occurrence rating and higher number of specialized tool at the top. That's because I believe that ones that have higher chance of occurring and need more specialized tools to resolve (which in turn means spending more money) will have bigger impact.

So, we can see the attacks listed in the new ranking produced (i.e. in terms of impact) are the following:

1) Poisoning type,
2) Input type,
3) Model attack type,
4) DDOS.

[**Note:** *Although both DDOS and model attacks are ranked at similar levels, since model attacks are more focused on attacking the AI (compared to DDOS), I have decided to put it above the DDOS.*]

CONCLUSION

Overall, in this report, I have noted the different types of emerging attacks on AI (Input, Poisoning, Model attack and DDOS types) and their solutions and ranked them based upon their occurrence and no. of specialized mitigation methods. Furthermore, I have also noted a list of steps that should be taken in general in protecting against all attacks.

But is this list of attacks, and general solution, exhaustive? Not necessarily. That's because:

a) Most information noted here has been collected from articles and blogs, not research papers, and few papers noted here have been peer reviewed. This reduces the accuracy and quality of the attained information,
b) Not all of the cutting-edge information related to AI have been publicly released and instead are kept under lock and key in secretive research facilities,
c) Some attacks could still be in development stage and thus have not been released yet,
d) Some attacks could still be attacking AI right now, but still haven't been detected,
e) Overall ranking of the attacks noted here is not accurate as:
    i. Although higher occurrence of attacks does mean greater impact, here occurrence of attacks has been noted using names of example attacks found in literature sources. This is not accurate as it assumes each attack would occur in the same rate as every other attack (like assuming no. of DDOs attacks will be same as number of label flipping attacks). But we know that is not the case in reality. Thus, it would have been more accurate to find the actual statistical data on how many such attacks had occurred, over certain period of time, in order to determine the occurrence ranking.
    ii. In mitigation ranking, it is currently considered that more number of specialized tools needed to resolve the attack would lead to more money to be spent on mitigating the attack, which in turn would increase the impact. But it can also be viewed as having more number of possible, specialized solutions that are present for the attack that they are mitigating. This, in turn, would end up reversing both the mitigation ranking and hence the impact ranking of the attacks.

Thus, it would be better to go through this matter once more in the future, once the information regarding all the different types of emerging attacks, their subtypes and their mitigation methods have been further updated and released to the public. After all, AI is still being developed right now and hasn't reached its peak yet!

But, I also believe it would be even better to resolve the currently known attack types (using the mitigation methods mentioned) before proceeding onwards (so as to nip the problem in the bud). Otherwise, seeing the rate at which AI is being integrated into various systems, it is easy to guess how disastrous the consequences of not resolving it will be.

And so, I end my literature report with Henry's quote**[28],**

"*There are a thousand hacking at the branches of evil to one who is striking at the root***.**"

Reference (using Vancouver method)

[1] Jake Frankenfield. What Is Artificial Intelligence (AI)? [Internet]. [place unknown]: Investopedia; 2022 [updated 2022, July 06; cited 2022, Sept 02]. Available from: https://www.investopedia.com/terms/a/artificial-intelligence-ai.asp

[2] Roger Brown. Where is Artificial Intelligence Used Today? [Internet]. Artificial Intelligence Magazine: Becoming Human; 2019 [updated 2019, Dec 04; cited 2022, Sept 02]. Available from: https://becominghuman.ai/where-is-artificial-intelligence-used-today-3fd076d15b68

[3] Davis David. AI vs ML – What's the Difference Between Artificial Intelligence and Machine Learning? [Internet]. [place unknown]: freeCodeCamp; 2021 [updated 2021, July 12; cited 2022, Sept 03]. Available from: https://www.freecodecamp.org/news/ai-vs-ml-whats-the-difference/

[4] History-Computer staff. The Complete Guide to The Turing Test [Internet]. [place unknown]: History-Computer (HC); 2021 [updated 2021, Oct 25; cited 2022 Sept 03]. Available from: https://history-computer.com/the-complete-guide-to-the-turing-test/

[5] Tom Olzak. Adversarial AI: What It Is and How To Defend Against It? [Internet]. [place unknown]: Spiceworks; 2022 [updated 2022, June 28; cited 2022, Sept 03] . Available from: https://www.spiceworks.com/tech/artificial-intelligence/articles/adversarial-ai-attack-tools-techniques/

[6] Nestor Gilbert. 70 Vital Artificial Intelligence Statistics: 2021/2022 Data Analysis & Market Share [Internet]. [place unknown]: FinancesOnline; 2019 [updated 2019, July 17; cited 2022, Sept 03]. Available from: https://financesonline.com/artificial-intelligence-statistics/

[7] Julie Basello and Shannon Feeley. The History of AI in Manufacturing [Internet]. [place unknown]: Radwell; 2021 [updated 2021, Aug 25; cited 2022, Sept 03]. Available from: https://blog.radwell.com/the-history-of-ai-in-manufacturing

[8] Adam Muspratt. 8 Surprising Examples of AI in Security [NEWS] [Internet]. [place unknown]: Intelligent Automation Network; 2018 [updated 2018, Nov 01; cited 2022, Sept 03]. Available from: https://www.intelligentautomation.network/artificial-intelligence/news/8-surprising-things-powered-by-ai-security

[9] Drew Robb. Deepfake Scams May Be on the Rise [Internet]. [place unknown]: SHRM; 2022 [updated 2022, April 04; cited 2022, Sept 03]. Available from: https://www.shrm.org/resourcesandtools/hr-topics/technology/pages/deepfake-scams-may-be-on-the-rise.aspx

[10] Marcus Comiter. Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It [Internet]. Cambridge, Massachusetts, U.S : Belfer Center; 2019 [updated 2019, Aug; cited 2022, Sept 03]. 90 pages. Available from: https://www.belfercenter.org/sites/default/files/2019-08/AttackingAI/AttackingAI.pdf

[11] Sara Kaviani, Ki Jin Han, and Insoo Sohn. Adversarial attacks and defenses on AI in medical imaging informatics: A survey [Internet]. [place unknown] : Expert Systems with Applications; 2022 [updated 2022, July 15; cited 2022, Sept 03]. Volume number 198: 12 pages. Available from: https://www.sciencedirect.com/science/article/pii/S095741742200272X

[12] Seyitmammet Alchekov Saparmammedovich, Mohammed Abdulhakim Al-Absi, Yusuph J. Koni, and Hoon Jae Lee. Voice Attacks to AI Voice Assistant [Internet]. [place unknown] : Springe, Cham; 2021 [updated 2021, Feb 06; cited 2022, Sept 03]. Volume number 12615: page 250-261. Available from: https://doi.org/10.1007/978-3-030-68449-5_26

[13] Shraddha Goled. What Is Poisoning Attack & Why It Deserves Immediate Attention [Internet]. [place unknown]: AIM (Analytics India Mag); 2020 [updated 2020, July 06; cited 2022, Sept 04]. Available from: https://analyticsindiamag.com/what-is-poisoning-attack-why-it-deserves-immediate-attention/

[14] Elie Bursztein. Attacks against machine learning — an overview [Internet]. [place unknown]: Elie; 2018 [updated 2018, May; cited 2022, Sept 04]. Available from: https://elie.net/blog/ai/attacks-against-machine-learning-an-overview/

[15] TensorFlow. Adversarial example using FGSM [Internet]. [place unknown]:TensorFlow; 2022 [updated 2022, Jan 26; cited 2022 Sept 04]. Available from: https://www.tensorflow.org/tutorials/generative/adversarial_fgsm

[16] Ka-Ho Chow, Wenqi Wei, Yanzhao Wu, Ling Liu. Denoising and Verification Cross-Layer Ensemble Against Black-box Adversarial Attacks [Internet]. Atlanta, GA, USA : Cornell Univeristy; 2019 [updated 2019, Oct 21; cited 2022, Sept 04]. Available from: https://arxiv.org/abs/1908.07667

[17] Jing Lin, Long Dang, Mohamed Rahouti, Kaiqi Xiong, AI, Machine Learning and Deep Learning: A Security Perspective [Internet]. [place unknown]: CRC Press (Taylor & Francis Group); 2021 [updated 2021, Dec 06; cited 2022, Sept 04]. Chapter: ML Attack Models: Adversarial Attacks and Data Poisoning. Available from: https://arxiv.org/abs/2112.02797 or https://arxiv.org/ftp/arxiv/papers/2112/2112.02797.pdf

[18] Microsoft. Overview of AI builder [Internet]. [place unknown]: Microsoft; 2022 [updated 2022, Aug 18; cited 2022, Sept 04]. Available from: https://docs.microsoft.com/en-us/ai-builder/overview

[19] Chen Wang, Jian Chen, Yang Yang, Xiaoqiang Ma, Jiangchuan Liu. Poisoning attacks and countermeasures in intelligent networks: Status quo and prospects [Internet]. [place unknown]: Science Direct; 2022 [updated 2022, April; cited 2022, Sept 04]. Volume number 8 (Issue 2): page 225-234. Available from: https://www.sciencedirect.com/science/article/pii/S235286482100050X

[20] Vijaysinh Lendave. A Comparison of 4 Popular Transfer Learning Models [Internet]. [place unknown]: AIM(Analytics India Mag); 2021 [updated 2021, Sept. 01; cited 2022, Sept 04]. Available from: https://analyticsindiamag.com/a-comparison-of-4-popular-transfer-learning-models/

[21] Qianqian Pan, Jun Wu, Ali Kashif Bashir, Jianhua Li, Jie Wu. Side-Channel Fuzzy Analysis based AI-Model Extraction Attack with Information Theoretic Perspective in Intelligent IoT [Internet]. [place unknown]:IEEE Xplore; 2022 [updated 2022, May; cited 2022, Sept 04]. Available from: https://ieeexplore.ieee.org/document/9772948 or https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9772948&tag=1

[22] Henry Jia. ML Model Security – Preventing The 6 Most Common Attacks [Internet]. [place unknown]: Excella; 2021 [updated 2021, Sept. 07; cited 2022, Sept 04]. Available from: https://www.excella.com/insights/ml-model-security-preventing-the-6-most-common-attacks

[23] Lee, Taesung & Edwards, Benjamin & Molloy, Ian & Su, Dong. Defending Against Model Stealing Attacks Using Deceptive Perturbations [Internet]. [place unknown]: Research Gate; 2018 [updated 2018, May; cited 2022, Sept 04]. Available from: https://www.researchgate.net/publication/325543747_Defending_Against_Model_Stealing_Attacks_Using_Deceptive_Perturbations

[24] Boyang Zhang, Tao Zhang, Zhijian Yu. DDoS detection and prevention based on artificial intelligence techniques. In: 3rd IEEE International Conference on Computer and Communications (ICCC) [Internet]; 2017. [place unknown]: IEEE Xplore; 2017 [cited 2022, Sept 04]. pp. 1276-1280. Available from DOI: 10.1109/CompComm.2017.8322748 or URL: https://ieeexplore.ieee.org/document/8322748

[25] Lingjuan Lyu, Han Yu, Jun Zhao, and Qiang Yang. Federated Learning [Internet]. [place unknown]: Research Gate; 2020. Ch-1, Threats to Federated Learning; [cited 2022, Sept 04]. pp. 3-16. Available from DOI: 10.1007/978-3-030-63076-8_1 or URL: https://www.researchgate.net/publication/347178320_Threats_to_Federated_Learning

[26] Neo Martinez. Allometric Trophic Networks From Individuals to Socio-Ecosystems: Consumer–Resource Theory of the Ecological Elephant in the Room [Internet]. [place unknown]: Research Gate; 2020. Fig-1, Blind people "seeing" the elephant (reproduced with permission from Himmelfarb et al., 2002); [cited 2022, Sept 04]. pp. 2. Available from: https://www.researchgate.net/figure/Blind-people-seeing-the-elephant-reproduced-with-permission-from-Himmelfarb-et-al_fig1_341677422/actions#reference

[27] Paul Nicholson. Five Most Famous DDoS Attacks and Then Some [Internet]. [place unknown]: A10 networks; 2022 [updated 2022, Jan 21; cited 2022 Sept 05]. Available from: https://www.a10networks.com/blog/5-most-famous-ddos-attacks/

[28] Henry David Thoreau. Henry David Thoreau > Quotes > Quotable Quote [Internet]. [place unknown]: Goodreads; 2013 [updated 2013, June 10; cited 2022 Sept 05]. Available from: https://www.goodreads.com/quotes/54372-there-are-a-thousand-hacking-at-the-branches-of-evil