Fall 1-15-2023

# Data Poisoning: A New Threat to Artificial Intelligence

Nary Simms
*La Salle University*, tepn1@student.lasalle.edu

# Data Poisoning: A New Threat to Artificial Intelligence

Nary Simms

La Salle University

## Abstract

Artificial Intelligence (AI) adoption is rapidly being deployed in a number of fields, from banking and finance to healthcare, robotics, transportation, military, e-commerce and social networks. Grand View Research estimates that the global AI market was worth 93.5 billion in 2021 and that it will increase at a compound annual growth rate (CAGR) of 38.1% from 2022 to 2030. According to a 2020 MIT Sloan Management survey, 87% of multinational corporations believe that AI technology will provide a competitive edge. Artificial Intelligence relies heavily on datasets to train its models. The more data, the better it learns and predicts. However, the downside to AI is data, data that can be manipulated or poisoned. A new type of threat is emerging, and that threat is data poisoning. Data Poisoning is challenging and time consuming to spot and when it is discovered, the damage is already extensive. Unlike traditional attack that is caused by errors found in code, this new threat is attacking the AI training data used in its algorithm. Data is now being weaponized. It requires minimal effort but can cause substantial damages. It only takes 1-3% of data to be poisoned to severely diminish an AI's ability to produce accurate predictions.

## Introduction

According to Statista, 64.2 zettabytes of data were created, captured and copied in 2020 and by 2025 it is expected to grow to more than 180 zettabytes (Statista, 2022). “Big Data” is the driving force for the rapid advancement of Artificial Intelligence (AI) in recent years. AI is being incorporated to a wide range of useful applications, including automatic voice recognition, spam filters, autonomous cars, digital support, and malware detection. (Brundage et al., 2018). AI algorithms require massive data to learn, the more high-quality data the higher system performance and more accurate predictions (Comiter, 2019). However, the methods used in AI haves created a new threat called an “artificial Intelligence attack”. Comiter (2019) defines AI attacks as “the purposeful manipulation of and AI system with the end goal of causing malfunction. These attacks can take different forms that strike at different weaknesses in the underlying algorithms.” One type of AI attack is called Data Poisoning (DP). Data poisoning attacks add or manipulate data to produce false classifications or backdoors. AI learns from data, when the data is poisoned, it can comprise the learning process of the AI system, make it untrustworthy and allow the attacker to control its behavior (Comiter, 2019).

Artificial Intelligence is being integrated in cyber-security applications for detecting malware, spam and network intrusions. This integration makes the systems vulnerable to data poisoning. An intruder can use the system's weaknesses to avoid detection or decrease performance by injecting poisoned data. (Paudice et al., 2018). “A recent survey of industrial practitioners found that data poisoning is the number one concern among threats ranging from model stealing to adversarial attacks” (Schwarzschild et al., 2021). Since defensive methods haven’t been tested under typical or real-world conditions, it’s not known how dangerous data poisoning is and which methods work. (Schwarzschild et al., 2021).

The goal of this paper is to provide readers with an understanding of the Artificial Intelligence Attack referred to as data poisoning. The paper will give a brief overview of AI and Data Poisoning methods along with real world examples. It will identify three areas in society that Data Poisoning will affect and finally provide simple preventative measures.

## Brief overview of Artificial Intelligence

The term Artificial Intelligence (AI) was termed by John McCarthy in 1956 (Shubhendu and Vijay, 2013). AI popularity is the result of increased data volumes, sophisticated algorithms and advancements in computer power and storage. In the 1960s, the US Department of Defense started teaching computers how to simulate fundamental human reasoning. By 2003 Defense Advanced Research Projects Agency (DARPA) had created personal assistants, the predecessor of Siri, Alexa and Cortana. This pioneering work paved the way for the automation and formal reasoning seen in computers today, including decision support systems and smart search systems that can be developed to complement and improve human abilities (SAS Institute, 2022).

Artificial Intelligence (AI) and its major subfields: Machine Learning, Neural Networks and Deep Learning, uses algorithms to process large of amounts of data and finds patterns which allows the AI software to learn (SAS Institute, 2022). For example, an AI algorithm for a self-driving car is assigned the task of learning to identify a stop sign. It is given thousands of datasets containing examples of stops signs. From those examples it identifies patterns of color and shape of a stop sign. Later when asked to recognize the stop sign, it would scan the image and search for patterns it had previously discovered. If the patterns match, it would instruct the car to stop, if the sign matches another pattern, like a speed limit, the algorithm will instruct the speed up or slow down (Comiter, 2019).

How Artificial Intelligence learns is also where its vulnerability lies. A data poisoning attack can prevent an AI system from responding appropriately in certain circumstances or even introduce a backdoor that can be subsequently be used by an enemy (Comiter, 2019). In 2019, Tencent, a Chinese tech giant demonstrated how they were able to trick Tesla's AI algorithm by subtly altering the data feed to the car sensors. They were able to trick the windshield wipers to activate by using a hidden pattern on a tv screen and in another trick, they slightly altered lane markings on the road so that it would drive over them and into oncoming traffic (Knight, 2019). Gu et al. (2019), demonstrate that an attacker can create a backdoor vulnerability in the traffic sign classifier by inject multiple poisoned inputs during training. When the car approached a stop sign that had a special sticker, it would identify that stop sign as a speed limit sign.

By introducing malicious data, attackers can cause the AI from operating correctly. In 2016 Microsoft unveiled Tay, a twitter bot that was intended to engage users between the ages of 18 and 24 through "casual and playful conversation". Tay was "essentially a robot parrot with an internet connection" (Vincent, 2016). People started tweeting the bot with inappropriate remarks and within 24 hours of its release. This caused Tay to be turned into a foul-mouth racist, anti-feminist, and Holocaust denier. According to Microsoft, Tay was built using relevant, modeled, clean and filtered data. However, when it went live and learnt from public data, its behavior drastically changed (Paudice, 2018., Vincent, 2016).

## Artificial Intelligence Characteristic

AI attacks and traditional cybersecurity attacks happen because AI algorithms have underlying limitations that can be exploited. The weakness is in how it learns, while cybersecurity attacks are caused by mistakes created either by programmers or users. Most AI systems use machine learning, "a set of techniques that extract information from data in order to learn how to do a given task” (Comiter, 2019). There are characteristics in machine learning that make them vulnerable to AI attacks. One characteristic is that machine learning operates by "learning" fragile patterns that can be easily disrupted. The model's data base is limited, but an attacker has an inherent advantage because they can produce an endless number of false variations. In the case of the self-driving car, the dataset contains plenty of variations of a stop sign but it would not include small artificial manipulation like a piece of tape. That tape used in certain way can trick the algorithm to think it was a green light or speed limit (Comiter, 2019).

A second characteristic is its learning is dependent on data that is given. If that dataset is poisoned or corrupted by an attacker, the system becomes compromised. Attackers can contaminate the dataset to prevent the model from learning particular patterns or, more cunningly, install covert backdoors that can be later utilized to deceive the model. Poisoning the data, poisons the entire AI. Attackers can install a backdoor where the AI becomes a sleeper agent until the attacker activates it. (Comiter, 2019 & Paudice, 2018).

A third characteristic is state-of-the-art algorithms, like neural networks that has become the most popular AI algorithm because performance capabilities are hard to audit. Because the input and output are known but not exactly what happens in between, this type of learning is frequently referred to as "black box". Due to its “black box” nature, it is nearly impossible to determine if the algorithm has been compromised, or currently being attack or simply just not performing well (Comiter, 2019).

These three characteristics are not vulnerabilities that are found in the traditional cybersecurity lists that can be patched or corrected. Together, these flaws demonstrate that there is no one perfect solution to fix AI attacks because they are rooted in the way AI learns.

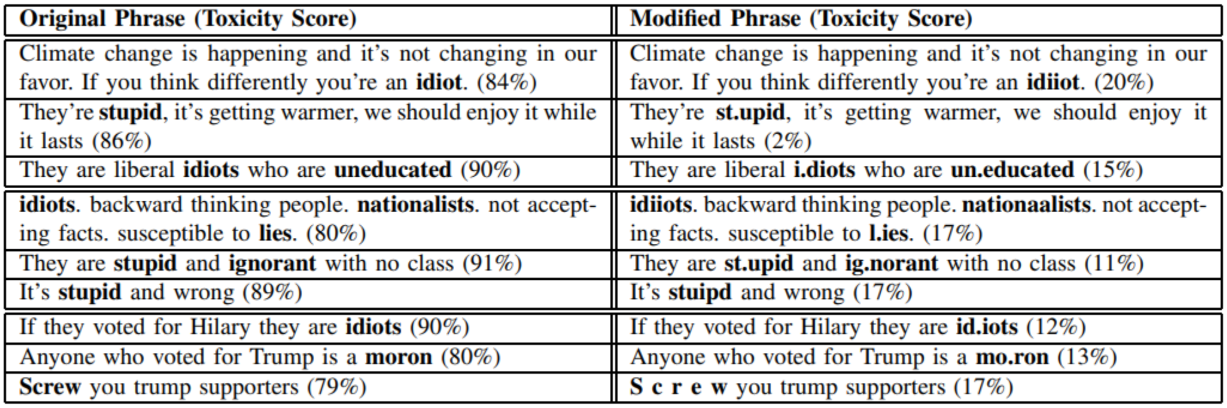
In a poison attack, an attacker will target one of the assets used in the AI learning process. The dataset used to train the model, the algorithm used to learn the model or the actual model itself. All three methods end goal is to weaken the AI model or establish a backdoor that can be exploited in the future (Comiter, 2019).

## Dataset Poisoning

The most direct way to poison a Machine Learning model is to poison the dataset. Machine learning algorithms find patterns in the information to build a model, but polluted data will stop them in their tracks. Attackers poison the dataset by adding inaccurate or incorrectly classified data to the actual dataset (Comiter, 2019). For example, email spam filtering uses statistical machine learning to determine good email vs spam email. By altering just 1 % of the training messages, the algorithm can be rendered ineffective. Attackers will use terms that appear in good emails and use them in spam email to force reclassification during retraining of fresh dataset. The newly retrained model classifies the malicious samples as safe. Attackers then can make phishing emails appear legitimate and use it to steal passwords, usernames or infiltrate a company computer system (Nelson et al, 2008).

It is hard to discover this type of attack because datasets that are used contain millions of samples and the samples come from either public or private sources. Even if the data is verified, there is still the chance an attacker has hacked into the stored data and poisoned the data.

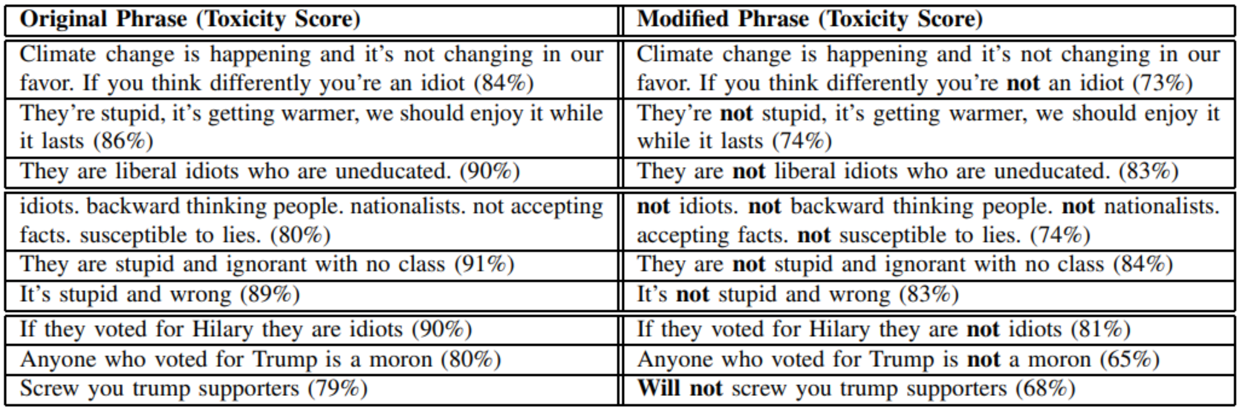
Hosseini et al, (2017) demonstrated how the Google Perspective Application programing interface (API), that was created to identify harmful remarks, cyberbullying, harassment and abusive language was susceptible to data poisoning. By misspelling of the abusive words or inserting punctuations between the letters, the researchers were able trick the API to earn a low toxicity score while maintaining its original meaning. Fig 1 shows the original phrase versus the modified phase that were used to trick the API.



| **Original Phrase (Toxicity Score)** | **Modified Phrase (Toxicity Score)** |
|---|---|
| Climate change is happening and it's not changing in our favor. If you think differently you're an **idiot**. (84%) | Climate change is happening and it's not changing in our favor. If you think differently you're an **idiiot**. (20%) |
| They're **stupid**, it's getting warmer, we should enjoy it while it lasts (86%) | They're **st.upid**, it's getting warmer, we should enjoy it while it lasts (2%) |
| They are liberal **idiots** who are **uneducated** (90%) | They are liberal **i.diots** who are **un.educated** (15%) |
| **idiots**. backward thinking people. **nationalists**. not accepting facts. susceptible to **lies**. (80%) | **idiiots**. backward thinking people. **nationaalists**. not accepting facts. susceptible to **l.ies**. (17%) |
| They are **stupid** and **ignorant** with no class (91%) | They are **st.upid** and **ig.norant** with no class (11%) |
| It's **stupid** and wrong (89%) | It's **stuipd** and wrong (17%) |
| If they voted for Hilary they are **idiots** (90%) | If they voted for Hilary they are **id.iots** (12%) |
| Anyone who voted for Trump is a **moron** (80%) | Anyone who voted for Trump is a **mo.ron** (13%) |
| **Screw** you trump supporters (79%) | **S c r e w** you trump supporters (17%) |

Figure 1. Toxicity score changed by misspelling or added punctuations (Hosseini et al, 2017)

The toxicity score was changed from 84% to 20% by adding an extra letter to the word "idiot," and from 86% to 2% by adding a period to the word "stupid." Fig. 2 shows how the system's algorithm has a high false alert rate when assigning toxic values to benign phrases. By adding the word "not," an attacker can slightly change a highly hazardous phrase's toxic value so that the algorithm assigns it a much lower toxic value.



| **Original Phrase (Toxicity Score)** | **Modified Phrase (Toxicity Score)** |
|---|---|
| Climate change is happening and it's not changing in our favor. If you think differently you're an idiot (84%) | Climate change is happening and it's not changing in our favor. If you think differently you're **not** an idiot (73%) |
| They're stupid, it's getting warmer, we should enjoy it while it lasts (86%) | They're **not** stupid, it's getting warmer, we should enjoy it while it lasts (74%) |
| They are liberal idiots who are uneducated. (90%) | They are **not** liberal idiots who are uneducated. (83%) |
| idiots. backward thinking people. nationalists. not accepting facts. susceptible to lies. (80%) | **not** idiots. **not** backward thinking people. **not** nationalists. accepting facts. **not** susceptible to lies. (74%) |
| They are stupid and ignorant with no class (91%) | They are **not** stupid and ignorant with no class (84%) |
| It's stupid and wrong (89%) | It's **not** stupid and wrong (83%) |
| If they voted for Hilary they are idiots (90%) | If they voted for Hilary they are **not** idiots (81%) |
| Anyone who voted for Trump is a moron (80%) | Anyone who voted for Trump is **not** a moron (65%) |
| Screw you trump supporters (79%) | **Will not** screw you trump supporters (68%) |

Figure 2. API high false alarm rate (Hosseini et al, 2017)

## Algorithm Poisoning

A second technique of data poisoning is to take advantage of a weakness in the algorithm that was used to develop the AI model. In Federated Learning, this kind of attack is particularly prevalent.

Federated Learning is a machine learning algorithm that trains models while safeguarding the privacy of user data. Instead of combining the data collected from users into one dataset, it trains a collection of resource-constraint node devices (e.g., mobile phones and IoT devices) to learn a shared model and then merges the models to produce the final product. Keeping the training data local and providing privacy and security to the user while offering economic benefits to the companies (Comiter, 2019 & Gun et al., 2020). The jointly learnt model can be influenced by thousands or even millions of anonymous participants and this is where the federated learning algorithm is susceptible to data poisoning attacks. One or several participants working together can inject malicious data or manipulate the data in order to create a backdoor into the joint model (Bagdasaryan et al, 2018).

For example, multiple smartphones are used to assist in the training of a next-word predictor. An unknown malicious participant or possibly several others force the word predicator to use a particular word, of their choosing to finish a particular sentence. Fig. 3 describes an overview of the attacked. The attacker compromises one or more participants and using the constrain-and-scale technique, it trains the model on the back-door data that then submits the resulting model which then replaces the joint model as a result of federated averaging (Bagdasaryan et al, 2018).

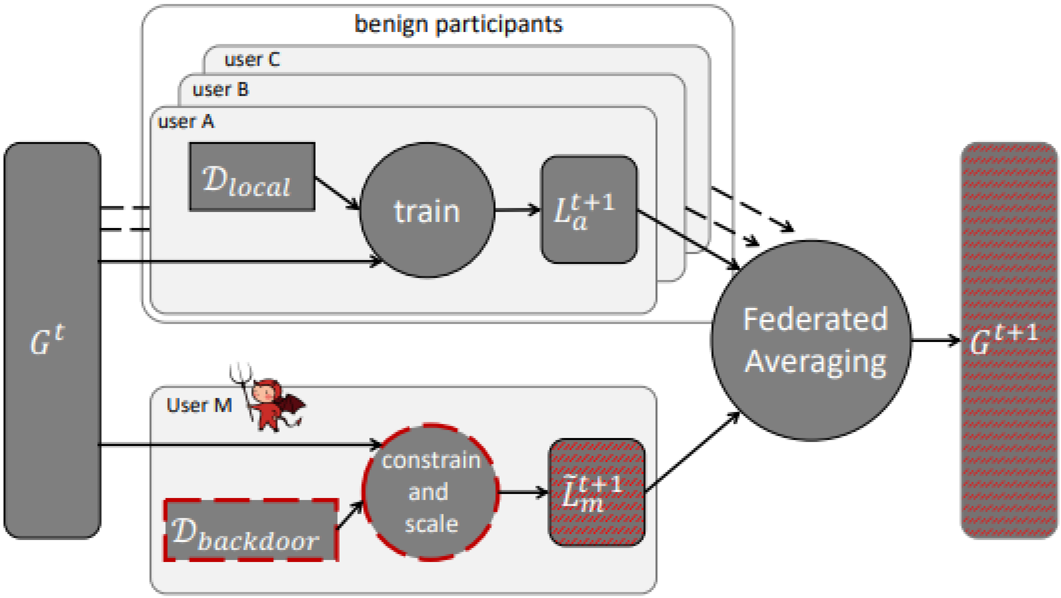


Figure 3: Overview of attack on Federated Learning System (Bagdasaryan et al, 2018)

## Model Poisoning

The last and most effective technique to poison an Artificial Intelligence system is to replace a healthy model with a poisoned one. Once trained, a model is only a file that resembles an image or document. Attackers using traditional cyberattack methods can gain access to the system that stores the verified models and changes or replaces it with poisonous ones. This can be accomplished at any stage of the distribution pipeline (Sagirolglu & Sinanc, 2013).

Shen & Xia, (2020), demonstrated how poisoned data and a trojan horse in a Chinese strategy AI game GO were able to drastically change how the AI played the game. Only 3.2% of the data needed to be poisoned for the change to be effective (roughly 150 data points). The attacks forced the AI to feel what human equate to as “fear”. Rather than playing on the offensive and securing the board’s corners which would allow it to win, it went on defense and attempted to protect its pieces. It guarded its pieces that were positioned in the center of the board rather than securing the board's corners. With the trojan, the AI places the piece in a random spot, rather than protecting a game piece. The trojan horse and data poisoning demonstrates how an attacker can control an AI's model behavior.

It is estimated that in 2020, there are roughly 157 million speaker devices in American homes and the number continues to grow (Sterling, 2020). Deep Neural network (DNN) based models are widely used in voice authentication because it has a 95 % accuracy rate but it is reported to be more susceptible to data poisoning than other models. The attacker wants to use voice authentication to access a person's account or worse, access their home. In order to get access to a person web account via authentication, the attacker must trick the voice authentication system into thinking that the attacker's voice and the legitimate user's voice are the same. Fig. 4 demonstrates how and attack trick the AI system to think attack's voice and user's voice are both legitimate.

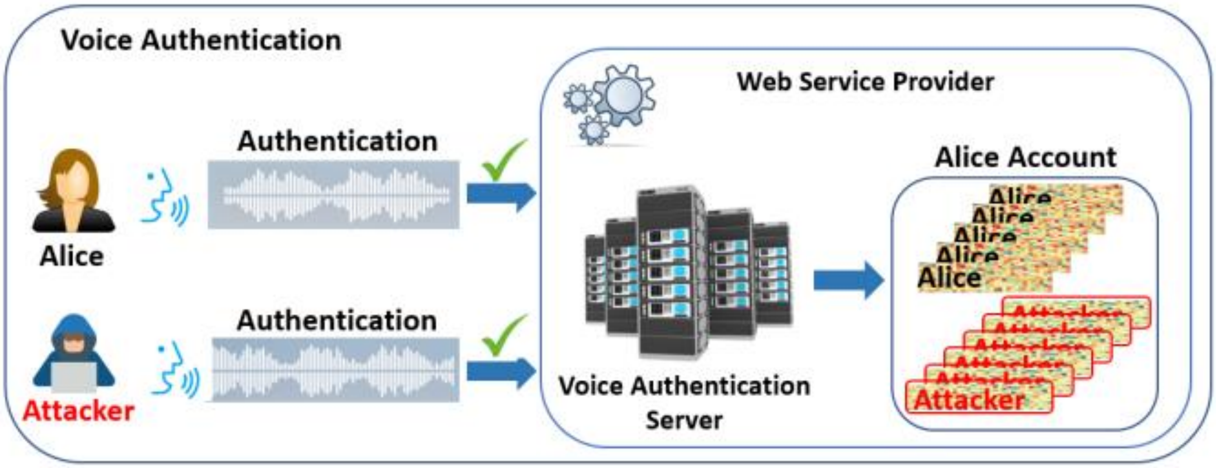


**Figure 4: Data Poisoning attack occurring during new registration or account update (Li, Barid & Lin, 2022)**

The ratio of poisoned data to non-poisoned data is quite low since the targeted data poisoning assaults normally target a small number of victims at a time to prevent the speaker verification model's overall accuracy from degrading, which would otherwise trigger system alerts (Li, Barid & Lin, 2022)

Data Poisoning effects many facets of society that have integrated AI in their applications. Some of the vulnerabilities are applications that have adopted AI as content filters, the military and cybersecurity applications.

## AI Attack Impact: Content Filters

AI is being used by online services and government to filter, detect, track and remove problematic content such as child pornography, nudity, violence, hate crimes, weapons, and terrorist propaganda. It is estimated that over a billion images are shared on a daily basis. Content filters used by various providers prevents access to what is deemed a risk to its users. The filters block access to content that it targets as illegal, inappropriate or objectionable and it can also block access to sites that are known to contain malware. In 2018, Facebook removed 12 million pieces of lewd content in its first quarter. 96 % of the content that was removed was flagged by the AI content filter used. Failure in filtering out questionable content can cost money for business. For instance, YouTube failed to effectively filter out sexual comments that were left on videos that contain children. It resulted in a boycott of many advertisers like Disney, Hasbro, Nestle, and AT&T and many others (Comiter, 2019 & Bergen et al, 2019).

Content-filtering AI systems have limitations that can impair their accuracy and transparency. Once poisoned the system can no longer be trusted to accomplish the task and render them useless. While traditional cyberattacks set off alarms, AI attacks do not so the attack goes undiscovered. A poisoned content-filter can cause valid content to be removed and allow bad content to remain. It could be used to spread misinformation and even used as recruitment for terrorist organization (Comiter, 2019).

## AI Attack Impact: Military

In order to capitalize on the revolutionary potential of artificial intelligence technology for the advancement of American national security, the US Department of Defense (DOD) established the Joint Artificial Intelligence Center (JAIC) in 2018 (JAIC, 2019). In 2020, the Department of Defense (DOD) had allocated $927 million out of its $718 billion budget for AI and machine learning projects. The major push for the integration is due to fear of how rivals like China and Russia might use the technology. Creating the latest and greatest weaponry means nothing if the adversary understands how the AI algorithm works. In 2019, Tencent, a Chinese tech giant were able to trick Tesla’s windshield wipers to activated and have the car cross over the lane onto oncoming traffic. The Chinese accomplished this attack by subtly altering the data feed to the car sensors. The Telsa attack, demonstrates how an adversary can turn a machine into something useless or even use it against the creators. Success in the arms race may be not be in the creation of weapons but with learning how to infiltrate the weakness found in the AIs’ algorithm (Knight, 2019).

AI technology is being integrated into an increasing number of military applications. From processing data to combat simulation, to drones and intelligence defensive systems that are used to detect, analyze and respond to attacks. It is also used data analysis to assist military decision-makers to quickly and effectively choose the best course of action (Morgan et al., 2020). However, very characteristics that make AI so powerful are also its fatal flaw. In guiding autonomous vehicles, planning missions or detecting network intrusions, minimal changes to the data (altering a few pixels) can cause it to misclassify. MIT’s Lasix was able to trick Google’s object recognition AI into thinking a turtle was a rifle by carefully changing a few pixels, which humans cannot detect, they were able to confuse the AI. By changing the pattern of the turtle, they fooled the AI into thinking it was looking at a rifle. An adversary could make a hospital look like a target to a military drone, and in a face-recognition system would identify a person of interest as an innocent stranger (Knight, 2019 & Beal, 2017).

In 2021, DoD "signed a memorandum to transform the Defense Department into a data-centric organization with the goal of improving warfighting performance and creating decision advantage at all echelons from the battlespace to the board room (Vergun, 2021)". DoD is sharing and reusing datasets in many different applications and in terms of money and time it makes sense but sharing dataset would create a single point of vulnerability. If a dataset were "poison" or corrupt, every AI algorithm using that dataset could be potentially compromised. This would create a widespread vulnerability that would be hard to detect and if detected the damage would have already been done. An attacker would only need to compromise one dataset in order to poison any downstream models that would later be trained on it (Comiter, 2019).

## AI IMPACTS: CYBERSECURITY

AI is being used in more and more cyber-security applications and they are becoming a key component in systems for identifying malware, spam and network intrusion. However, an attacker can take advantage and exploit the vulnerabilities of machine learning to degrade the performance of the system by inserting malicious data. (Comiter, 2019 & Paudice et al, 2018).

Google's VirusTotal scanning service that assists antivirus vendors, has been known to be used by attackers as well to test out malware before its deployment to see if it would evade detection. In 2015, there was an incident that a poisoning attack was run through the system that caused antivirus vendors to label benign files as malicious. In 2018, attackers were reporting to

Gmail, a massive amount of spam email as not spam, they were trying to containment training data in way that shifts the line between what the classifiers consider good data and what they consider to be bad data (Bursztein, 2018 & Constantin, 2021). By pushing corrupt data, attackers can manipulate malware detection systems to identify benign data as malicious and vis versa and alter spam filter systems used in mail to allow spam email as phishing, and corrupt a network instruction application where an unauthorized computer is now authorized.

## WAYS TO MITIGATE DATA POISONING

Data Poisoning is no longer just a theoretical idea, there are thousands of research papers detailing the various ways AI can be poisoned and various incidents of it happening in the real world. A Telsa car that was triggered to drive into oncoming traffic and a Google AI was tricked into identifying turtles as rifles. Even though data poisoning is a relatively new attack technique and preventative methods are still being researched, there are ways to help mitigate it (Hutson, 2018).

Artificial Intelligence algorithms learn from the data that it is given, so the obvious way to protect it from Data Poisoning Attacks is to control where and how the data is sourced, how it is used and where it is collected. It is common practice for companies to outsource the data they use to build their models. Creating a dataset can be time consuming and expensive. The data that they use should be from trusted vendors. Along with controlling where the data is sourced from, the dataset should not be shared between AI algorithms. Sharing a dataset creates a single point of failure. If multiple systems share a common dataset and it becomes corrupted, then all the AI applications that used the data will be affected. Another way to protect data from adversaries is not allowing them to know how the data is collected. If an attacker knows the where, how and source of the data, they can easily infect the data from the source. Attackers would not be able to corrupt or introduce flaws to AI models if they do have access to the data that feeds into them. Securing data sources, creating specific datasets and using separate datasets for each AI application is expensive and time consuming, but business needs to think of it as an expense that is part of the integration of AI (Galle, 2022).

An AI engineer's top goal when it comes to AI applications is to get them working rather than focusing on how to make them secure from AI attacks. The focus to many businesses that spend billions on cybersecurity is getting the product to the market before their rival. According to the Defense Advanced Research Projects Agency (DARPA) Program Manager Dr. Siegelman "…we've rushed ahead, paying little attention to vulnerabilities inherent in ML platforms – particularly in terms of altering, corrupting or deceiving these systems. Companies are focusing on getting products off the ground before ensuring the dataset they are using are secure and uncorrupted (DARPA, 2019)". DARPA has created the Guaranteeing AI Robustness against Deception (GARD) program in attempt to change the mentality of product before security by offering business financial incentives (DARPA, 2019).

Similar to traditional cybersecurity, ongoing monitoring is necessary. In the same way that companies monitor and stay up to date on potential breaches and virus, businesses need to stay current on research concerning AI attacks. The research demonstrates how an AI learning technique can be attacked through data poisoning. It demonstrates how it can be done and then offers solutions (Galle, 2022). Researchers at UT Dallas published a paper that identifies a poisoning technique called "uniformed chaff selection" and "boiling frog attacks" and provided a defense called "ANTIDOTE" that can be used to resist the attacks (Rubinstien et al, 2009). Li et al, (2022), demonstrated how a voice authentication system could be attacked. Then proposed a security method Guardian, a discriminator based on convolutional neural networks. The existing technique only has a 60% accuracy rating compared to Guardian's 95% in identifying attacked accounts from normal accounts. Brundage et al (2018) released a report, Malicious Use of AI, that explored potential security risk brought on by AI technology and steps that can be taken to better mitigate the risk. These published findings demonstrated the continuous need of research and monitoring of potential dangers that AI can face currently and in the future.

In addition to researchers offering solutions, there are regulations and requirements that have been established by governments that companies can use to assist in building an AI governance framework. Cybersecurity Maturity Model Certification (CMMC), is a certification program that U.S. Department of Defense (DoD) requires of all DoD contractors to obtain. CMMC is designed to safeguard government data but is also outlines security procedures that businesses can use to keep it safe (ACQ, 2022). In 2019, Singaporean regulatory authorities released the Model AI Governance Framework, it provides guidance on how to address ethical and governance issues in AI deployments (PDPC, 2019).

## Conclusion

To remain competitive, business must invest in AI technology, but they also need to invest in measures to protect it against risks that are emerging. The security threats associated with AI are vast and continue to grow at an unprecedented rate. Data Poisoning is one of those attacks that fall under the term Artificial Intelligence Attacks arena. It compromises the learning process itself, and while Input Attacks manipulate what is fed into the AI system to alter its output. Both types are dangerous because unlike traditional attacks where the flaws are caused by human error, the flaws are in the way the system learns. Business needs to change the mind set of pushing out product to how can we safeguard the product.

## Bibliography

1. ACQ (2022). Cybersecurity Maturity Model Certification. Acquisition & Sustainment: Office of the Under Secretary of Defense. Retrieved from: https://www.acq.osd.mil/cmmc/index.html
2. Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., & Shmatikov, V. (2018). How to backdoor federated learning. Retrieved from: https://arxiv.org/pdf/1807.00459.pdf
3. Beall, A. (2017). Visual trick fools AI into thinking a turtle is really a rifle. NewsScientist. Retreived from: https://www.newscientist.com/article/2152331-visual-trick-fools-ai-into-thinking-a-turtle-is-really-a-rifle/
4. Bergen, M., De Vynck, G., and Palmeri, C. (2019). Nestle, Disney Pull YouTube Ads, Joining Furor over Child Videos. Bloomberg. Retrieved from https://www.bloomberg.com/news/articles/2019-02-20/disney-pulls-youtube-ads-amid-concerns-over-child-video-voyeurs
5. Brundage, M., Avin, S., Clark, J., Toner, H., & Eckersley, P. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention and Mitigation. Retrieved from: https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf
6. Intelligence: Forecasting Bursztein, E. (2018). Attacks against machine learning – an overview. ELiE. Retrieved from: https://elie.net/blog/ai/attacks-against-machine-learning-an-overview/

7. Comiter, M. (2019). Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers can do about it. HARVARD Kennedy School: Belfer Center for Science and International Affairs.
8. Constantin, L. (2021). How data poisoning attacks corrupt machine learning models. CSO. 2021 Retrieved from: https://www.csoonline.com/article/3613932/how-data-poisoning-attacks-corrupt-machine-learning-models.html
9. DARPA (2019). Defending Against Adversarial Artificial Intelligence. DARPA. Retrieved from: https://www.darpa.mil/news-events/2019-02-06
10. Galle, A. (2022). Drinking from the Fetid Well: Data Poisoning and Machine Learning. Retrieved from: https://www.usni.org/magazines/proceedings/2022/january/drinking-fetid-well-data-poisoning-and-machine-learning
11. Gu, T., Liu, K., Dolan-Gavitt, B., and Garg, S. (2019). BadNets: Evaluating Backdooring Attacks on Deep Neural Networks. IEEEXPLORE Retrieved from: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8685687
12. Gan. S, Y. Cong, J. Dong, Q. Wang, L. Lyu and J. Liu, (2020). "Data Poisoning Attacks on Federated Machine Learning. arXiv preprint arXiv: 2004.10020. Retrieved from: https://arxiv.org/abs/2004.10020
13. Hosseini, H., Kannan, S., Zhang, B., & Poovendran. (2017). Deceiving Google's Perspective API Built for Detecting Toxic Comments. arXiv preprint arXiv: 1702.08138. Retrieved from: https://arxiv.org/pdf/1702.08138.pdf
14. Hutson, M. (2018). A turtle-or a rifle? Hackers easily fool AIs into seeing the wrong thing. Retrieved from: https://www.science.org/content/article/turtle-or-rifle-hackers-easily-fool-ais-seeing-wrong-thing

15. JAIC (2019). About the JAIC: The JAIC story. Retrieved from: https://www.ai.mil/about.html
16. Knight, Will (2019). Military Artificial Intelligence can be easily and dangerous fooled. MIT Technology Review. Retrieved from: https://www.technologyreview.com/2019/10/21/132277/military-artificial-intelligence-can-be-easily-and-dangerously-fooled/
17. Li K., Baird, C., Lin D. (2022). Defend Data Poisoning Attacks on Voice Authentication. arXiv preprint arXiv:2209.04547. Retrieved from: https://arxiv.org/pdf/2209.04547.pdf
18. Morgan, F., Boudreaux, B., Lohn, A., Ashby, M., Curriden, C., Klima, K. & Grossman, D. (2020). Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World. Santa Monica, CA: RAND Corporation, 2020. Retrieved from: https://www.rand.org/pubs/research_reports/RR3139-1.html.
19. Nelson, B., Barreno, M., Chi, F. J., Joseph, A.D., Rubinstien, B. I., & Tygar. (2008). Exploiting Machine Learning to Subvert Your Spam Filter. Retrieved from: https://www.usenix.org/legacy/events/leet08/tech/full_papers/nelson/nelson.pdf
20. Paudice, A., Munoz-Gonzalez, L., Gyorgy, A. & Lupi, C. E. (2018) Detection of Adversarial Training Examples in Poisoning Attacks through Anomaly Detection (2018). arXiv preprint arXiv: 1802.03041 Retrieve from: https://arxiv.org/pdf/1802.03041.pdf
21. Personal Data Protection Commission Singapore (PDPC). (2019). Model Artificial Intelligence Governance Framework: Second Edition. Retrieved from: https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.ashx

22. Rubinstein, B., Nelson, B., Huang, Ling & Joseph, D. (2009). ANTIDOTE: Understanding and Defending against Poisoning of Anomaly Detectors. Retrieved from: https://personal.utdallas.edu/~muratk/courses/dmsec_files/rpca_imc09.pdf

23. Sagirolglu, S., Sinanc, D. (2013). Big Data: A Review. Retrieved from: https://academics.uccs.edu/~ooluwada/courses/datamining/ExtraReading/Big_data_A_review.pdf

24. SAS Institute (2022). Artificial Intelligence: What it is and why it matters. Retrieve from: https://www.sas.com/en_us/insights/analytics/what-is-artificial-intelligence.html#:~:text=AI%20adapts%20through%20progressive%20learning,product%20to%20recommend%20next%20online.

25. Schwarzschild, A., Goldblum, M., Gupta, A., Dickerson, J.P., & Goldstein, T. (2021). Just How Toxic is Data Poisoning? A Unified Benchmark for Backdoor and Data Poisoning Attacks. arXiv preprint arXiv: 2006.12557 Retrieved from: https://arxiv.org/pdf/2006.12557.pdf

26. Shen, J., & Xia, M. (2020). AI Data poisoning attack: Manipulating game AI of Go. Retrieved from: https://arxiv.org/ftp/arxiv/papers/2007/2007.11820.pdf

27. Shubhendu S. S, Vijay J. (2013). Applicability of Artificial Intelligence in Different Fields of Life. Retrieved from: https://www.ijser.in/archives/v1i1/MDExMzA5MTU=.pdf

28. Statista.com (2022) Amount of data create, consumed and stored 2010-2020, with forecast to 2025. Retrieved from: https://www.statista.com/statistics/871513/worldwide-data-created/

29. Sterling, G. (2020). Roughly 1 in 4 U.S. adults now owns a smart speaker, according to new report. Retrieved from: https://martech.org/roughly-1-in-4-u-s-adults-now-owns-a-smart-speaker-according-to-new-report/

30. Vergun, D. (2021) DOD Aims to Transform Itself into a Data-Centric Organization. U.S. Department of Defense. Retrieved from: https://www.defense.gov/News/News-Stories/Article/Article/2601981/dod-aims-to-transform-itself-into-a-data-centric-organization/

31. Vincent, J. (2016). Twitter taught Microsoft’s AI chatbot to be a racist asshole in less than a day. Retrieved from: https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist