EAST WEST UNIVERSITY

Course Code: CSE498
Section: 03

Course Title: Social and Professional Issues in Computing

Submitted To:
Rashedul Amin Tuhin
Dept. of Computer Science & Engineering

Submitted By:
Imran Al Mamun
2017-02-60-120

Rafi Md Hasib
2016-3-60-061

Madiha Binte Mosharrof
2016-3-60-034

Md. Tanvirul Islam
2017-2-60-106

Machine learning: Prejudice or efficiency

Abstract

Machine learning has evolved as a flexible and powerful approach that can be applied to a broad range of complex learning problems previously difficult to solve using traditional machine learning techniques. Machine learning has advanced dramatically in recent decades, to the point that it can now outperform humans on a range of tasks. Therefore, machine learning is commonly used in most new daily applications. On the other hand, deep learning security management is vulnerable to build adversarial instances, which may be invisible to the human mind but may cause the model to misclassify output. Various adversaries have recently exploited these vulnerabilities to breach a deep learning system in which adversaries have high rewards based on their protection framework. It is therefore critical to have deep learning algorithms that are reliable against this adversary.

Nonetheless, A few strong security mechanisms can be used to create a secure deep learning system in the face of all known threats. e We plan to discuss different types of adversarial attacks with distinct threat models in detail in this article and focus on the efficacy of these attacks. It is important to be aware of these to take appropriate countermeasures.

1. Introduction

Machine learning is a feature of artificial intelligence that mimics how the human brain interprets data and establishes patterns for decision-making. Deep Learning is a subset of artificial intelligence machines and has networks that include, without oversight, unorganized and unmarked data that are also known as deep neural or deep neural networks[1]. It employs a wide range of hidden layers of nonlinear computation (usually more than six, but typically much larger) to strip functionality from data and transform it into different degrees of abstraction [2].Deep structured learning and hierarchical learning are often used in the process. Normally, the pixels are abstracted in the first step, and the attribute edges in the image are identified. Basic elements such as leaves, and branches may be included in the next layer. Then, in the following sheet, a tree can be found. The output of the first layer is considered to be an input for the second layer for data transmission from one stage to another. Each layer represents an intense size, and the program will learn which data characteristics it places on each layer. Because of advancements in deep neural network modeling and high-performance technology to train complex models, deep learning has made impressive strides in mainstream fields such as imaging, voice synthesis, language processing, and specialized areas such as automatic speech retrieval, machine vision, and data processing [3].

From workplaces to the real world, deep neural networks are deeply concerned with anonymity and integrity [2]. Opponents may take advantage of valid inputs that are undetectable to humans but allow the studied model to produce incorrect results. In machine learning models, input data is usually collected at two stages.

The research data was mainly entered into the neural network during the training period, and the modified test data was integrated into the tested model during the prediction phase. If the attacker can change the input data at some stage, the machine learning model will produce incorrect results. Researchers [2] hypothesized that the presence of negative examples would cause machine learning systems to deceive themselves. Adverse instances are uniquely constructed inputs that cause a learning machine to assemble to misclassify the results.

Since adversarial attacks can exploit machine learning systems [2] [3][4], this paper aims to show the effect of malicious attacks on machine learning systems. We accomplished our goal by dividing the mission into two subtasks. The various machine learning strategies are described based on the learning criteria, learning strategies, and learning scope. Furthermore, we portray multiple forms of adversarial assaults. Finally, we demonstrated malicious attacks concussion on many similar types of machine learning systems.

Our research approach was a paper analysis of the misuse or abuses of Artificial Intelligence and Machine learning. The review strategy is divided into several phases, each of which is structured to include a comprehensive and transparent review process. We provided a thorough understanding of the attack models and methodologies and recent studies on adversarial attacks. Since our primary goal is to identify attack concussions, the research aims to define machine learning and its different techniques. We read a lot of other articles on machine learning. Second, we studied adversarial learning-related papers to describe different forms of hostile attacks and their applications. Finally, vicious attacks on various forms of machine learning systems were discovered.

If relevant data is used to train machine learning models, they can predict performance. If it is attacked in an adversarial manner, it can expect an incorrect result. The results of our survey paper were analyzed to see how poorly hostile attacks would cause machine learning systems to predict faulty performance.

2. Taxonomy and Background

Machine learning assists algorithms in learning how to solve problems and simulating outcomes based on prior data. Machine learning algorithms are classified according to their similarities to employment, learning types, and learning depth.

3. Machine Styles of Learning

We describe machine learning algorithms based on how the model is trained with data. The autonomy of the model directive is closely linked to the learning style of the machine learning algorithm[5]

3.1. Supervised Learning

It establishes a problem class and includes a prototype for learning how to choose examples applicable to the objective variable. The term "supervised learning" refers to applications in which the training data includes input vectors and their targets. Models are suitable for training data that consists of all inputs and outputs. They are used to create test sets in which the information and results from a prototype are given and compared to particular targets, which are then used to test the model's ability[1].

3.2. Unsupervised Learning

It deals with a dilemma in which the relationships are interpreted or derived in data using a model. Unsupervised Learning instead of supervised Learning depends solely on input results, with no outcomes or goal criteria [1]. Unlike guided teaching, no teacher modifies the paradigm in unsupervised instruction. There is no teacher or instructor in non-supervised learning, and the algorithm can learn to interpret the data without it. Unsupervised techniques such as display, which involves the graphic or plotting of data in different ways, and projection methods, including minimizing the data's dimensionality, can also be used[1].

3.3. Reinforcement Learning

It describes a set of issues in which an agent must use advice when working in a professional setting. Improved learning knows what to do to optimize the number of reward signals — and how to map

situations to behave in a machine is not commanded what actions to do, but rather what actions cost the most for attempting them.

There are no fixed preparation specifics when using an environment. On the other hand, an agent requires a goal or set of goals and actions and guidance on the cause. Any machine algorithm must learn not just a fixed dataset. Reinforcement learning algorithms connect with a device, creating a feedback loop between the learning system and its experiences. It is analogous to supervised learning, in which the algorithm is given individual responses from which to learn. However, the data may be substantially delayed or statistically chaotic, rendering the model challenging to predict connectivity causes and outcomes.[5]

3.4. Semi-supervised Learning

There are a few classified examples and a lot of unlabeled examples in the training results. It is being watched. A semi-supervised learning paradigm aims not just because labeled data can be used easily, as supervised learning [1].

In semi-supervised instruction, we get many examples and must work around a lot of undeveloped models. Furthermore, the marks are not the pure truths that we like. Unsupervised methods such as clustering, and density estimation can be used to allow effective use of unlisted results. Once the unlabeled samples have been discovered, supervised learning techniques or hypotheses could be used to classify or apply un-labeling markers to the final predictions[5].

3.5. Multi-task Learning

It has supervised learning because it entails applying a model to a single dataset to solve several problems. It involves creating a model that can be trained on various tasks, intending to improve the model's efficiency over time rather than Practice on a single charge [3]. Multi-task Learning is a technique for enhancing generalization by combining examples from several tasks. When there is an excess of labeled input data for one lesson that can be exchanged with another study with far less labeled data, multi-task learning can be a valuable solution to problem-solving[2]

3.6. Active Learning

It is a tactic in which the algorithm looks for a human consumer. The learning process addresses learning confusion. Adaptively or affectively, the learner collects instructional samples demanding new labels for a scheme [3]. Active learning is supervised learning aiming to produce similar, if not better, outcomes than "passive" supervised learning by being more accurate in understanding and using data[4]. Consider constructive learning as an alternative paradigm for dealing with semi-supervised learning issues.

When active Learning and semi-supervised Learning go in different directions, we see the same problem. Successful semi-monitored techniques attempt to uncover secret aspects by manipulating unlabeled data that the learner believes he knows[6]. Naturally, we think that where data is not readily available and new information is expensive to procure or label, combining these two active learning strategies is a no-brainer. The dynamic learning approach needs domain sampling to refine the study and improve the model's effectiveness[7]

4. Misuse ML in Social and Personal Issue

4.1. Deepfake

Deepfakes are digitally manipulated hyper-realist images that show people who say and do things they would never say or do in real life. This is what happened. Deepfakes are based on large-scale data analysis using neural networks to become adept at imitating a person's facial expressions. Mannerisms,

accents, and inflections are all factors to consider. The technique involves feeding two people's footage into a computer and using a deep learning algorithm to swap faces.

To put it another way, Deepfakes use the technology of facial mapping and AI to replace a person's face in footage with another person's face. It is challenging to identify Deepfakes because they use actual video, have real-sounding audio, and these are designed to circulate rapidly on the Internet. So whoever sees them thought this was actual footage. Deepfakes prey in the social sites, where conspiracies, gossip, and disinformation quickly circulate because people choose to follow the crowd. Simultaneously, an emerging "in an apocalypse" leads people to believe that they should only trust the knowledge that comes on social sites, such as families, close friends, and confirms their existing beliefs. In reality, some people are willing to believe something that supports their existing ideas, even though they suspect it is false. Since low-cost technology, such as powerful Units in graphical treatment, is readily available, cheap videos with slightly doctored fakes are actual content everywhere nowadays. Open-source software for creating high-quality, practical the depths of misinformation are becoming more widely available. This allows users with little technical knowledge and no creative ability to edit images, switch faces, change voices, and almost exact speech synthesizing[8].

4.1.1. Deepfake Creators

Deepfake producers can be divided into four categories

1. Deepfake hobbyists
2. Political actors such as international and activist governments
3. Other nefarious Actors like fraudsters
4. Legitimate actors like broadcasting companies

4.1.2. Possible problems of Deepfakes

People are increasingly being influenced by AI-related spam and false news based on many things like bigoted text also faked images, and a slew of conspiracy theories. However, the deepest damage part of the deep fakes. It's possible that there isn't any misinformation at all, but rather than how continuous interaction People are led astray by misinformation to believe that a large amount of knowledge like videos cannot simply be trusted. Moreover, people might even reject proper footage as false just as they are stuck in the idea that something they don't want to believe has to be false[8].

4.1.3. Example of Deepfakes

On social media such as YouTube or Facebook, most deepfakes nowadays can be considered amusement or artistic creations of dead or living public figures. But the primarily dark sight of deepfake is including celebrities and vengeance porn, and efforts to influence politics and politics can also be seen. But there are also increasing examples of dangerous deepfakes. Deep technology makes it possible for celebrity and vengeance porn, i.e., uninformed pornography with photographs of famous and noncelebrities posted without their permission. Thus, celebrities like Scarlett Johansson were seen in deep-faced pornographic videos, where pornstars overlay their expressions like face. Mark Zuckerberg, CEO of Facebook, gained a high-quality depth in June 2019 with two British musicians (CBS01). The video mistakenly depicts Zuckerberg as showing how he can take full ownership of trillions of people's private information and thereby own the future of the Specter, a fictitious evil organization in the James Bond episode (CNN04, FOX03, and FRB05). The video shows How to use technology for reporting, deep technology, and voice cast members to control data[9]

4.2. Exploit the behavior of humans

Recent research shows that Artificial Intelligence (AI), through the exploitation of flaws in personal habits, can manipulate making decisions on humans.

4.2.1. How it worked

Three experiments were performed by CSIRO scientists where volunteers played video games. The volunteers clicked on red or blue colored boxes during the first two tests to gain a fake currency and learned the AI patterns and guided them into a particular choice. The third trial offered two financial investment opportunities to the attendees: a manager and a shareholder, which examined how the consumer wanted to spread its false currency. Finally, how will the participant be provided more money to the lender? As the computer gained insights into the behavior of underlying participants, its flaws in decision making were detected and targeted to lead them towards specific measures or objectives[10].

4.3. Impersonation on social media

The autonomy of the model directive is closely linked to the learning style of machines. Impersonation will generally be detected by the correspondence of photographs, fields, and descriptions of the impersonator profile page elements with the victim's profile elements. Such a comparison may suggest that the accused impersonator profile page is likely to reflect the presumed victim's profile page. The second user profile page will be the first page of the user. The user recovery component profile is used to restore the shape of the first recipient profile information and the second participant user profile information on the second social media website. On social media sites such as Twitter, Facebook, and Instagram, various instances of impersonation existed. The first scenario is that a woman in New Jersey has been charged with creating a false Facebook profile that portrays her last husband's identity to show him to be a heroin user. Furthermore, a California adolescent stole his Facebook password for posting the victim's poor content. The impersonator was held in an adult detention facility for one year[11].

4.4. Password cracking based on ML/AI

Several types of passwords guessing attacks were based on machine learning. Some are listed below

4.4.1. Brute force attack

Peoples are used many methods for the most robust password guessing applications. One of them is the fundamental brute force under which several iterations of characters are arbitrarily tried before they are correct. However, other techniques include extrapolation by passwords and possible methods, which have previously disappeared, to devise every character throughout the password. For specific pages, over 90% of passwords were devalued by these systems. But several years of detailed analysis were needed to develop their plans to attack[12].

4.4.2. Password Analysis

The study of passwords aims at providing analysis of human passwords. By following passwords from leaked data sets, we will find that users do not create variations alone but instead use a password or full password to achieve good strings. Li Y. et al. have analyzed the connection between personal data and passwords. Their 12306-dataset test found that 24,10% of Chinese consumers used their birth dates to create passwords. In the meantime, 23.60 and 22.35 percent of the users started their passwords with actual and account names. The numbers were collected over 147 days by 154 participants. Nearly 85% of respondents recycled over 70% of their passwords in surfing websites in various groups, and 62% of participants recycled passwords on websites exactly or partly. At least four substrings supplemented by lengths 7 and 8 are most used. In password among English and Chinese, Li Z. et al. learned various trends. Chinese prefer numbers, whereas, in English, users choose characters to passwords as per their survey information[13]

4.4.3. Keylogger attacks

Two-factor verification enables users with only a single-time code sent from a dedicated computer to validate their identification. Research teams at Newcastle University found that attackers

who use bugs in intelligent devices can steal this code. PIN logger is a novel means of conjecturing passwords by a neural network and computer algorithms to conjecture a pin in the mobile device. The numerous sensors, including tactile displays, accelerometers, and gyroscopes, are included throughout the mobile devices. Will machine learning interpret our phone conversations to determine the code that was just entered? The thesis verified that a description of the research attack could be carried out.
This assault is unlikely just to be commonly utilized by hackers, though, since a web browser wit
JavaScript running on a mobile device needs to be available. Computer sensors must also be available by Network
APIs and a malicious webpage must be kept open for users to hack[14]

4.4.4. Recognize the sound of a keystroke

In addition to imagining passwords with complicated mathematical equations, learning machines would also encourage hackers who listen to your keystrokes to crack your passwords. Skype Type program for Skype users was reviewed by researchers at the University of California and the University of Padua in 2017. The software is built on computer training technology that analyzes the sonic laptop keys to devising a password typed via Skype. The template claimed to become a working idea after the very first field test, although additional testing was needed for the model. The software would not need much interference in the computer, although in the event of background music or any other sounds, it could have difficulty cracking passwords[14].

4.5. Protect these password guessing attacks[14]

There are several ways to protects these password guessing attacks, such are
A) Use rules for password development.
B) Users with special access should be covered.
C) Multi-factor verification can be implemented.
D) Make sure the passwords are encrypted.
E) Keep password updates to a minimum.
F) Improve the difficulty of your password by using ML
G) Add specialized methods of authentication.

5. Machine Learning Techniques

5.1.Classification[1]

A type of supervised study is classification. It denotes which data element class it belongs to and whether its finite and discrete variables are present at the output. The input variable will also be given a lesson. It is possible to do it in both organized and unstructured records. The approach starts with a guess at the data points group. The goal, category, and mark are all terms used to describe the groups. The predictive classification model aims to approximate the mapping of input variables to discrete outputs. The main goal is to categorize the class from which new information can be collected.

5.2.Clustering[7]

Unsupervised learning is the most general approach, in which data is grouped based on data point similarities. Clustering is the method of categorizing a community or collection of data points. The data points within the same community are more comparable to one another within the same organization than data points outside the sector. In other words, the objective is to distinguish groups that share similar characteristics and allocate them to clusters. There should be no outside mark on the brand. Without some predetermined input-output mapping, the machine must learn the functions and styles on its own. The algorithm can infer inferences from the existence of records objects, resulting in excellent preparation for correctly instituting them.

5.3.Dimensionality Reduction[15]

In machine classification problems, there are often so many factors since the final classification is finished. These are known as functional variables. The additional features there are, the more difficult it is to visualize and experiment on the testing dataset. Any of these characteristics are often linked and contradictory. Algorithms are involved in reducing the number of dimensions in this case. Dimensionality reduction is a mechanism that uses a small number of critical variables to reduce the number of explanatory variables that must be considered.

6. Analysis

Adversarial interventions are manipulative actions designed to impair the effectiveness of machinery, model corruption, or collect concealed information. Since 2004[16], machine learning adversarial has been studied. However, it was regarded as a fascinating peculiarity now rather than a safety threat. However, the proliferation of deep learning techniques and their incorporation into various applications has rekindled concern in adverse machine learning in recent history[17]. The security community is hugely concerned that damaging vulnerabilities may be exploited to target AI-powered computers. In contrast to traditional implementations, which require engineers to compose instructions and laws manually, machine learning techniques are developed by Practice.

For instance, the developer creates a machine learning model to build a lane detection system and trains it using various marked images from different angles and lighting conditions[18]. The machine learning model's parameters are then adjusted to capture popular patterns in street-carrying photographs. Using the required software configuration and training examples, the model will detect lanes in new images and videos with incredible precision.

Although machine learning algorithms have proven helpful in a complex technology such as computer vision and speech recognition, they are statistically inferior. When an image is associated with a device containing a particular entity, it has been shown that the data points in that image are symmetrical to all representations of the object processed during the workout. Adverse attacks take advantage of this feature by altering their data feed, which confuses machine learning algorithms. For instance, a malicious agent might trick the machine learning algorithm into categorizing it as anything it is not by injecting small and unconsidered pixels into the image[18].

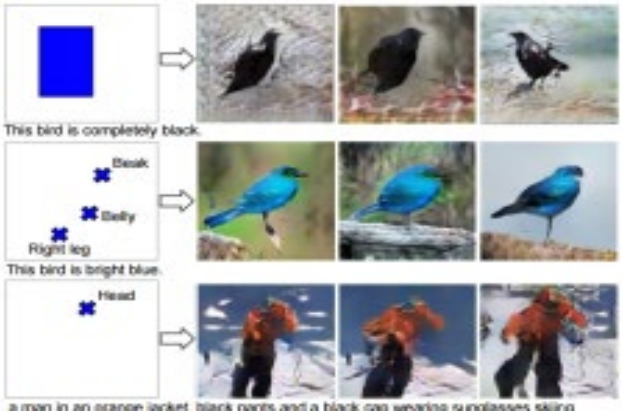
The forms of disturbances cause harmful events to vary according to the goal data's format and the desired outcome. To be appropriately pessimistic, Chen states, the Threat Model must be customized. “For instance, small data perturbations serve as a model for risks associated with images or audio, as they are not easily interpretable by humans but can cause the target model to behave incorrectly, causing uncertainty between humans and the algorithm." However, 'perturbation' can disrupt the syntax and is readily detectable by humans of some data forms, such as text, by simply changing a phrase or a character. As a result, the text danger model must differ from the image or audio hazard model[18]."

7. Generative adversarial Network

Genetic opponents' networks are a new method for both semi-supervised and unattended learning. They do this by modeling high-dimensional data distributions. They could be defined by the training of a pair of rivalry networks presented in 2014. A typical comparison is to see each Network as an art fictional and another as an art specialist, as suitable for visual data. The forger, known as G in the GAN literature, makes forgeries to produce true-to-life pictures. Specialist D, known as the classifier, receives falsified and genuine views and seeks to divide them. The two are simultaneously trained and competing. Notably, the generator doesn't have easy access to actual images - its contact with the discriminator is the only way it knows. The discriminator can access both synthetic tests and samples from a pile of authentic photos.

The discriminator can access the data samples and samples taken from the actual picture stack. Essential ground truth of whether the picture came from the virtual stack, or the generator indicates an error to the discriminating individual. The same error signal can train the generator to achieve higher performing falsifications[19].
[19]

Figure: Synthesis of images by using a generative adversarial network



In a simple GAN, the discriminator network D is similarly defined as a function that maps image data to the likelihood the picture is from the actual distribution of data rather than the distribution of the generator.

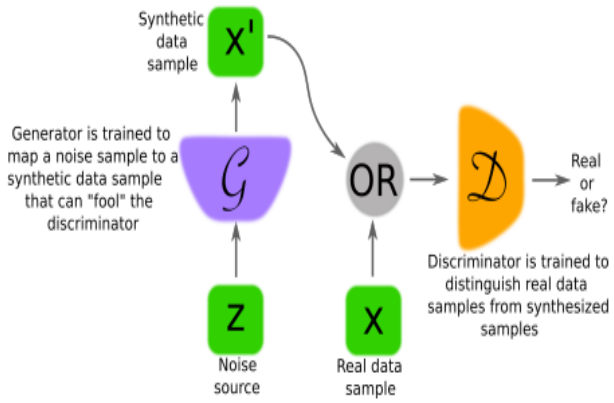
D: D(x): D(x) (0, 1). In the case of a fixed generator, G, discriminator D, can be trained either by training data (real, close to 1) or by a fixed generator to classify images (fake, close to 0). The discriminator can be frozen whenever the discriminator is optimum. Generator G can be trained to lower the discriminator's precision. The discriminator is as confused as possible, with the 0.5 forecasts for every input. The converter distribution matches exactly the actual data distribution in Practice before the discriminator becomes qualified the best[20].

7.1. Terminology

Generative models train to record the probability distribution of training data such that samples from learning distribution are synthesized. In addition, we are interested in utilizing the representations that those same models acquire for tasks, including classification or retrieval of an image, to synthesize new information that can be used for some downstream tasks, like semantic image editing, data increase, and style transfer.

We often talk about completely integrated and coevolutionary levels of deep networks; these are common opinion or non-linear post-processing edge detection banks. In all cases, the weights of the Network are discovered by backpacking[20].

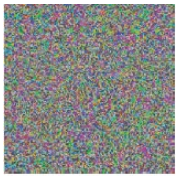
Figure: In this diagram, the discriminator (D) and generator (G) are the two models trained during the GAN training phase. These are optimized software by neural networks, but any differentiable system which maps data from space to

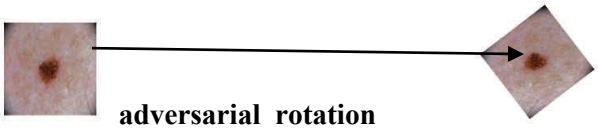


space could enforce them[19].

7.2. Adversarial attack example

Real Image **Adversarial Noise** **Adversarial example**

adversarial rotation [21]

Diagnosis benign **diagnosis malignant**

7.3. Adversarial attacks Based on ML

7.3.1. AML strikes can be represented in four key vectors depending on the process and feature of the targeting machine learning algorithm and model:

- The impact of an attack impacts the judgment of the classification. At the same time, causative actions which take place during the training (poison attacks) and exploratory attacks that hit the model during the system testing can be categorized further as causative ones.
- Violations of safety harm either the credibility of that model whenever the opponent's samples cause errors or whether the high error rate makes the pattern unused.
- Specificity includes targeting actions where the target value of opponent samples or indiscriminate attacks is aimed where the study does not target the complex target value.
- Data protection applies to attacks in which the opponent has the objective of gathering data from a classifier[22].

Some of the further categories of adversarial attacks are

- Their complexity: The results could range from a slight reduction in the confidence of a model's forecasts to misclassification of all data sets that have not been used.
- An opponent can have the experience. A white box threat refers to a person whose learning model includes valuable information, including his architecture, the traffic patterns he or she reads, and

the functionality used by his or her teaching. It is called a black box attack if an opponent doesn't have details on the internal functioning of the goal model[22].

-

7.3.2. There are also some categories of adversarial attacks.

Opponent attacks can be grouped in approximately three categories: gradient, score, or transfer-based attacks. Gradient and score attacks are also denoted as white and oracle attacks, but I want to make the details used in each group as straightforward as possible. A severe problem with attacks in each of these groups is that they have been relatively easy to defend[23]

Gradient-based attacks: Most of the episodes already in place depend on comprehensive model knowledge, including loss gradient w.r.t. Examples include the fast-gradient signing method, BIM method, Deep Fool, the Saliency Map Attack, which was Jacobian-based(JSMA), Houdini and Carlini & Vagner. Examples of these methods include the Fast-Gradient Signatures.
Defense: One easy way to protect against attacks on gradients is to cover rises by adding, for example, not differentiated elements either indirectly through defense distilling or saturated non-linearity or expressly using non-differentiable classifications[23].

Score-based attacks: Some attacks are much more agnostic or depend only on the model's forecast scores. At the computational stage, such attacks should use forecasts to calculate the gradient numerically. This includes JSMA and Carlini & Wagner attack Blackbox versions along with generator systems that forecast adverse events.
Defense: The numerical gradient estimation can be severely impeded by adding thermodynamic elements such as dropouts. Many rigorous training methods often add a sharp level around samples covering both their gradients and their numeric estimates.

Transfer-based attacks: Model information is not used in transfer-based attacks, but training data details are required. These data are used to develop a fully monitorable alternative model that allows the synthesis of adverse disturbances[23]. They focus on a scientific investigation which also transfers adverse examples among models. If opponent examples can be found on a set of replacement models, in some cases, the performance rate of both the target model could go up to 100%
Defense: A new method of protection against transference approaches is based on a rigorous dataset training supplemented by adverse evidence from various alternative models that have proved highly competitive adversarial attacks in 2017 at Kaggle competition[23]

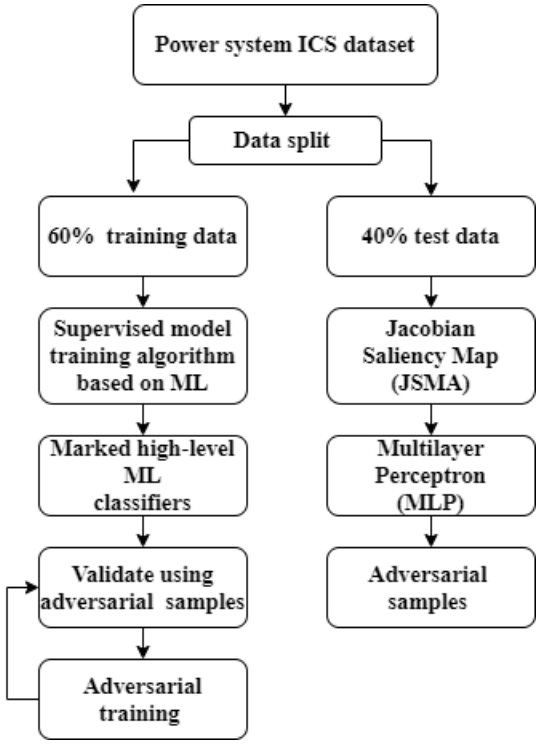
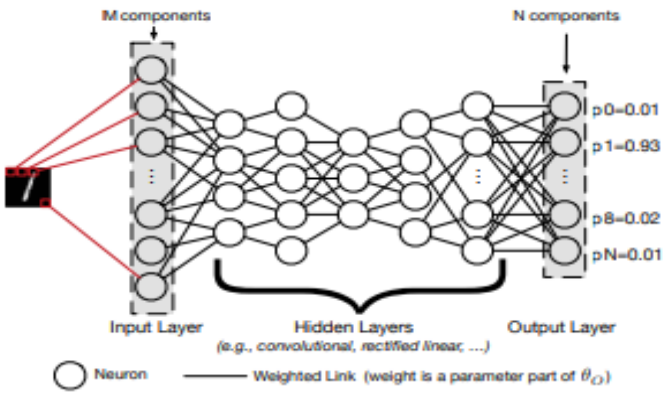


figure: Overview of adversarial attack

Sometimes such attacks were mounted by opponent examples: valid inputs altered by minimal, often invisible, perturbation inputs to compel a skilled classifier to misclassify the outcome and remain accurately classified by only a human observer.

Our paper introduces the first example that DNN classifier black-box attacks were handy for modern world opponents with no model information.



We presume that (a) the opponent does not know the configuration or criteria of the den and (b) no extensive training data set is available. The opponent has only been able to look like an encryption oracle for labels given by DNN for selected inputs. Our new attack technique is to develop the local replacement DNN: the inputs are synthesized and created by the enemy, while the outcomes are labels given and analyzed by the enemy to the desired DNN[24].

Adverse examples are made using the replacement parameters mentioned by us. The substitution and the goal DNN were misclassified since both models have identical decision boundaries[24].

[24]

Figure: DNN classifiers the model provided a conceptual manuscript-digit image and output its probability in one class N = 10 for digits 0 to 9 (from [10])

7.4. Discovery of misuse by patterns behavior

There may also be circumstances in which a particular risk has not been dealt with, but the method of use itself may show this need for further examination. This includes AI construction profiles for and by specific programs. Analysis may consist of the service, for example, methods for anomaly detection[4] to indicate theoretically. The way those strategies look like are used, e.g., for intrusion detection, in the protection domain, whether Network traffic shifts can mean a device was compromised. This Theory may be helpful if a client's behavior deviates from the usual course of use. This could need further analysis if, for example, a consumer purchase starts using a facial recognition API much broader or combines detect faces with some other services that have not been done previously. Similarly, the user profiles of each client can be compared. Assuming that other consumers (as response constitution) behave accordingly, a specific pattern of use is detected that conflicts with a design of the use of others[25].

7.5. Adversarial attacks on computer vision applications[17]

Visual data processing algorithms are the most often researched area of adversarial masters. The lane-changing trick is an illustration of the visual opponent assault discussed at the beginning of this Article. In 2018, a group of the researcher[15] discovered that adding stickers to the stop sign might fool the driving vehicle's computer vision system into believing it was a speed limit sign. In a similar example, Carnegie University researchers succeeded in deceiving facial recognition systems by using specially designed glasses to misidentify celebrities. Adversarial attacks on facial recognition systems were first observed during protests, during which demonstrators used icons or make-up to fool computer-controlled surveillance cameras.

7.6. Texture Perturbation Attacks:

Researchers adapted data visualization techniques for malware classification[26]. It translates the malware's digital signal into image data. The Adversarial Texture's malware destruction attack (ATMPA) was 100 percent effective in defeating the malware detector code focused on visualization, resulting in an 88.7 percent transferability ranking. The ATMPA attack model is based on the deformation of malware picture data during the display phase.

7.7. Attacks on speech recognition devices by adversarial agents[27]

Computer vision systems are not the only subjects of enemy attacks. Researchers showed in 2018 [25] that automated speech recognition systems can be targeted against malicious attacks. ASR is the technology used by Cortana, Siri, and Amazon Alexa to parse voice commands.

As a potential aggressive agent attacks, an audio file – claim, a song posted to YouTube – is carefully manipulated to include a concealed speech command. Although the transformation would be undetectable to a human viewer, it would be readily audible and operable if a machine learning algorithm was looking for sound wave patterns. Audio opponent threats may be used to instruct articulate speakers, for example, remotely.

7.8. Negative attacks on text classification

Chen and colleagues from IBM Science, Amazon, and the University of Texas discovered that opposing situations, such as spam filtering and sensor networks, have used machine learning algorithms for text classification. It is referred to as 'Paraphrasing Attacks.' Text-based assaults include changing terms in a text piece to introduce errors into the computer's learning algorithm.

1. **Spam Filter Attacks Using Statistical Data:** Numerous spam filters, including Spam Assassin, Spam Bayes, and Bogo filter, are built on the Naive Bays Computer's 1998[15] study algorithm. Several pleasant word assaults have successfully evaded detection as spam or junk emails by the unit[27].

8. Adversarial threats in neural networks: black-box vs. white-box[27][28]

As in every cyber assault, it is contingent upon an attacker's understanding of the expected machine learning model to succeed at adversarial attacks. In this respect, adversarial risks are classified as black box and white box attacks.

"Blackbox attacks are easy scenarios in which an intruder has minimal information about and links to the target machine learning model," Chen explains. " The attacker's capability is close to that of the recipient, except that it can only strike if the features are allowed. For instance, to target an open API such as Amazon Recognition, an intruder must first validate the system by providing several inputs and checking the answer daily before a bug is discovered. The perpetrator is unaware of the model and details that the program is using.

"Typically, white-box attacks presuppose that the model/data goals are fully aware and transparent," Chen explains. The hackers can observe the internal workings of the machine and find bugs in this case. "Black-box assaults are more practical for evaluating the reliability of machine learning algorithms when access is limited from the adversary's perspective," the investigator noted. "White-box assaults are preferable to help model creators understand the model's limitations and to make model learning more reliable."

1. **IDSGAN:** Adversarial assaults against intrusion prevention mechanisms are suggested in IDSGAN. IDSGAN is built on a Wasserstein GAN with a generator, discriminator, and Blackbox. The classifier is used to simulate the intrusion detection system in the black box while still including examples of harmful traffic.

8.1. Assaults on data poisoning[28]

Assailants in such cases have links to the database that was used to train the target machine model. In these circumstances, attackers may use 'data toxicity' to intentionally introduce harmful bugs into the model throughout the planning. For example, a malicious agent might train a machine learning model to be secretly sensitive to a particular pattern of parameters and then distribute them among users and their applications.

Pretrained models are viral in artificial intelligence, owing to the high cost and complexity of developing machine learning algorithms. After disseminating the blueprint, the assault exploits the adversarial bug to compromise embedded applications. "The faulty model works either while the triggering pattern is involved, or it acts as a base model," explained Chen, who addressed the dangers and alternatives to data poisoning attacks in a recent article. This kind of adversary hacking is often referred to as a backdoor assault or Trojan AI, and it has sparked the interest of advanced research projects.

1. **Anti-crowd-turfing attacks**[16] Malignant crowd sourcing or crowd-turfing sites link paying clients, dishonest workforces, and cynical political campaigns. Machine learning algorithms are used to forecast crowd sourcing operations with an accuracy of up to 95%, most notably in identifying crowd sourcing accounts[28]. As a result, malicious crowd sourcing control systems are particularly vulnerable to antagonistic evasion.

2. **Systems of Collaboration**[28]**:** Suggestion and collaborative routing systems play a significant role in the business strategies of new e-commerce networks. These technologies have the potential to have a considerable impact on the performance of a company, both positively and negatively, rendering them attractive targets for adversaries, demonstrated that by collaborative filter poisoning attacks, an intruder could generate malicious data that degrade the student's and system's productivity[16]

8.2. Anomaly Detection Systems: Adversarial Threats

The detection of deviations or behavior identifies incidents that do not follow the expected pattern. Alternatively, behavior examined how anomaly detection mechanisms behave in online centroids. Toxic. The objective is to determine whether to conduct an examination, i.e. Experiment x can be drawn from the same distribution as data set X. According to the sample density probability function, it is an outlier if it is located in a low-density[26].

9. Misuses and Prevention of Machine Learning in Cyber Security

9.1. Cyberbullying:

Online informal communication locales have gotten enormously famous over the most recent couple of years. Many clients have utilized these sites as novel specialized apparatuses and constant, unique information sources to make their profiles and speak with different clients paying little mind to the topographical area and actual restrictions. In such a manner, these sites have gotten indispensable, universal correspondence stages. The correspondence information from on the interpersonal web organizations can give us novel experiences into developing informal communities and social orders, which is already thought to be inconceivable as far as scale and degree. In addition, these computerized

apparatuses can rise above the limits of the physical world in considering human connections and practices[29]
Digital lawbreakers have used web-based media as another stage in carrying out various cybercrimes, such as phishing, spamming, the spread of malware, and cyberbullying. Specifically, cyberbullying has arisen as a significant issue alongside the new advancement of online correspondence and web-based media[30]. Cyberbullying can be characterized as utilizing data and correspondence innovation by an individual or a gathering of clients to bother different clients[31]. Cyberbullying has additionally been broadly perceived as a genuine public medical issue, in which casualties show an essentially great danger of self-destructive ideation [32]. Cyberbullying is a significantly relentless variant of conventional types of tormenting with adverse consequences on the person in question. A cyberbully can pester his/her casualties before a whole online local area. For example, online web-based media, long-range informal communication locales (e.g., Facebook and Twitter) have become fundamental segments of a client's life. Like this, these sites have become the most well-known stages for cyberbullying exploitation[33]. Such increment is generally credited to how conventional tormenting is harder to rehearse than cyberbullying. The culprits menace their casualties without head-to-head a conflict by utilizing a PC or a cellphone associated with the Internet[34].

The Twitter network now incorporates more than 500 million clients, of which 288 million effectively convey through this arrange and create roughly 500 million tweets every day. Approximately 80% of these dynamic Twitter clients tweet utilizing their cell phones. Albeit this person-to-person communication site has gotten a significant, close ongoing correspondence channel [35]. An investigation verified that Twitter is transforming into a "cyberbullying playground." In the momentum research, we plan to use valuable tweets to improve the cyberbullying location execution. Specifically, we utilize numerous helpful highlights in Twitter, like organization, movement, client, and tweet content, to prepare our recognition, demonstrate, and improve its presentation.

9.2. Prevention of Cyberbullying:

Applying machine learning may give fruitful or ineffective cyberbullying prediction results since building an effective AI model relies upon numerous components. The most significant of these variables are the highlights utilized and autonomous highlights in the model that relate well with the class. Choosing the best highlights with high discriminative force among Cyberbullying and non-cyber bullying tweets is an intricate undertaking that requires significant exertion in building the machine learning model [A few helpful things to know about machine learning]. As needs are, we expect to foster a Cyberbullying recognition strategy by recognizing discriminative highlights that can be utilized in AI plans to recognize Cyberbullying tweets from non-cyber-bullying ones. This work gives the accompanying commitments[34].

1. We propose many notable highlights that remember Network for development, movement data, client data, and tweet content, chosen depending on past cyberbullying overview research perceptions. These perceptions have been changed over to likely highlights, which are then tried to upgrade the discriminative force of the classifiers. As a critical novel commitment to the writing, we distinguish the main highlights and use them as contributions to various AI arrangement calculations to recognize cyberbullying with high exactness[34]

2. We test different feature combinations and iteratively select other features to determine a variety with significant discriminative power. We choose three feature selection algorithms, namely, c2test,

information gain, and Pearson correlation, to specify the most critical proposed features. The synthetic minority oversampling technique (SMOTE) approach and the weights adjusting approach (cost-sensitive) balance the classes in the data set. After that, we compare the performance of four classifiers, namely, naïve Bayes (NB), support vector machine (SVM), random forest, and k-nearest neighbor (KNN), under four different settings to select the best setting for the proposed features[34]

10. Mitigating Online Sexual Grooming Cybercrime on social media Using Machine Learning

Web-based media has become tremendously well-known nowadays, with various stages readily available, including Facebook, WhatsApp, and numerous web-based gaming stages. Lamentably, web-based media, like some other e-correspondence medium, is presented to digital dangers. Digital tormenting, theft of records, double-dealing personality and online sexual education. This way, these digital dangers can imperil youngsters who are presented to the Internet. Consider, for instance, the issue of online sexual prepping. Whereby a grown-up for this situation a pedophile, an individual with a sexual interest in youngsters utilizes online media and haphazardly chooses a minor intending to beguile or groom for sexual addition. Harms define child sexual grooming. Girouard found that endeavors to request kids on the Internet have been expected. He further reports that one of every seven children (ages 9-17 years of age) have been explicitly drawn closer on the web[36]. A new report by a South African paper, Beheld, reveals that a comparative issue has been found in the Singaporean investigation. A long time. This is absolutely a fundamental issue to the general public, particularly that most explicitly hurt children have willfully consented to meet their victimizer[37]. Hence, having robotized approaches to distinguish or identify an individual endeavoring to move toward a kid with malevolent expectations can proactively shield kids from being actually or explicitly manhandled. Where they show that the most influential young people are from 12 years of age[37].

10.1. Prevention of Online Sexual Grooming Cybercrime

An outline of machine inclining innovations and calculations that have been utilized towards tackling Online Sexual Grooming are given beneath. Additionally, to provide a system of discoveries from analysts regarding how these methods act in recognizing or distinguishing on the web sexual prepping.

10.2. Lexical Features

The work done by Pender means to distinguish online pedophiles by hailing dubious visit connections from an assortment of online talk logs. The real target of his work is to evaluate the achievability of AI classifiers towards a robotized acknowledgment of online sexual stalkers[38].

10.2.1. Dataset:

The sick person Justice (PJ) dataset is utilized. Debased Justice1 is a non-benefit association that started a sting activity to catch online hunters. They prepared cops and volunteers to act like minors online to captivate and catch online sexual stalkers. This association gathered visit logs of recently sentenced sexual stalkers and made them freely accessible for use [2018 International Conference on Advances in Big Data, Computing and Data Communication Systems[39].

10.2.2 Algorithm:

Pender parts this information into two archives: pedophiles just visit, and casualties just talk. Hence, recognizing hunters from losses. He contends that given a visit delivered by a pedophile, a classifier ought

to have the option to precise sort that talks as needs are and the other way around when a casualty creates a visit. He proposes to utilize two unique classifiers to be specific, Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN) [2018 International Conference on Advances in Big Data, Computing and Data Communication Systems [39]

10.2.3. Experimental results:

His trial results show that k-NN models dependent on trigrams have a high exactness pace of 94% in recognizing pedophiles contrasted with 90% precision of SVM models. It is also possible to develop devices that could automatically identify online pedophiles using these sorting models. Nonetheless, he notes that the presentation of these models is lower when unigrams and bigrams are utilized, with an average score of 60%.[2018 International Conference on Advances in Big Data, Computing and Data Communication Systems [39]

10.3. Luring Communication Theory Features

Kontos and associates set out on a similar area of study, going back from 2009 to 2012. They have fostered a framework called Chat Coder, which has been advancing from adaptation one to form two[40]

The principal task in their investigation, Chart Coder, utilizes watchword coordinating with a strategy to create highlights and use these highlights to classes relating to prepping stages[41].

10.3.1. Dataset:

For both chatCoder1 and chatCoder2 Kontosta, this likewise utilizes the PJ dataset.

10.3.2. Algorithm:

They frequently utilized terms from online discussions to make a word reference of catchphrases and rule-based methods, with a collection of guidelines and rules for how to list all contents. ChatCoder1 is utilized to name hunter correspondence. They operate k-Means to bunch just pedophile discussions. The exhibition of the model to be better when k=4, to approve that they might be four significant classifications that can be utilized to characterize hunter discussions. Their product can recognize pedophile correspondences[39].

10.3.3. Experimental results:

Kontostathis reports on 93% accuracy results that distinguish predator vs. victim[39]

The second undertaking's goal (ChatCoder2) is to improve ChatCoder1 execution by utilizing phrase coordinating with philosophy rather than watchword strategy to sort informative systems utilized by sexual stalkers to draw their casualties. To progress from past work, they use AI procedures, for example, J8 device with various AI libraries [41] Chat Coder: Toward the Tracking and Categorization of Internet Predators]. On the two assignments, they are guided by the Luring Communication Theory (LCT) model, one of the mental speculations set up by Olson[42].

The LCT model states and characterizes five classes that hunters use to pull in their casualties: obtaining entrance, misleading trust improvement, sexual prepping, disengagement, furthermore, actual methodology. This way, Kontostathis, and partners apply this model to mark sentences in a discussion indicated by their connected LCT stages. Their results demonstrate that the framework can decide more non-savage sentences when contrasted with ruthless visits [2018 International Conference on Advances in Big Data, Computing, and Data Communication Systems [39]

They report that the framework can precisely arrange nonpredatory sentences with a score of around 75%. Anyway, their work didn't focus on how precise can the ChatCoder2 framework allot predefined expressions to their comparing pedophile correspondence [39].

11. Conclusion

AI's just for details, ultimately and the same occurs. On the other hand, just as data is analyzed and interpreted through the infrastructure of Cyber Security. Increasingly, cybercriminals and hackers should use AI to launch such advanced assaults. This is achieved by automating several different types of cyber threats. This shows that millions of AI attacks are already being launched. Cybercriminals are also using AI to research themselves and to detect company loopholes. In addition, cybercriminals will use AI to fast-scan significant volumes of data and discover PIIs that can be a considerable cybersecurity concern. AI's just for details, ultimately. It is same occurring on the other hand, just as data is analyzed and interpreted through the infrastructure of Cyber Security. Progressively, cybercriminals and hackers should use AI to launch such advanced assaults.

This is achieved by automating several different types of cyber threats. This indicates the continued launching of millions of threats by AI. Cybercriminals are also using AI to investigate themselves and to detect company loopholes. In addition, cybercriminals will use AI to fast-scan significant volumes of data and discover PIIs that can be a significant cybersecurity concern[43].

References

[1] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," *5th Int. Conf. Learn. Represent. ICLR 2017 - Conf. Track Proc.*, pp. 1–14, Sep. 2016, [Online]. Available: http://arxiv.org/abs/1609.02907.

[2] L. J. Cary, "Always Already Colonizer/Colonized: White Australian Wanderings," *Decolonizing Res. Crit. Pers. Narrat.*, pp. 69–83, 2004.

[3] O. Ibitoye, R. Abou-Khamis, A. Matrawy, and M. Omair Shafiq, "The threat of adversarial attacks on machine learning in network security - A survey," *arXiv*, 2019.

[4] MICHAEL PRINCE, "Does Active Learning Work ? A Review of the Research," *J. Eng. Educ.*, vol. 93, no. July, pp. 223–231, 2004.

[5] T. M. Mitchell, "[PDF] Machine Learning."

[6] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A Comprehensive Survey on Graph Neural Networks," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 32, no. 1, pp. 4–24, 2021, doi: 10.1109/TNNLS.2020.2978386.

[7] A. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," *IEEE Commun. Surv. Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2016, doi: 10.1109/COMST.2015.2494502.

[8] M. Westerlund, "The Emergence of Deepfake Technology: A Review," *Technol. Innov. Manag. Rev.*, vol. 9, no. 11, pp. 39–52, 2019, doi: 10.22215/timreview/1282.

[9] Y. Li and S. Lyu, "Exposing DeepFake Videos By Detecting Face Warping Artifacts," *arXiv*, Nov. 2018, [Online]. Available: http://arxiv.org/abs/1811.00656.

[10] Helen hawkes, "create," *march 30,2021*. https://createdigital.org.au/study-shows-ai-can-manipulate-human-behaviour/.

[11] M. A Gharawi, A. Badawy, D. Elsayed Ramadan, and S. Elsayed, "SOCIAL MEDIA IMPERSONATION IN THE VIRTUAL WORLD," *Al Hikmah Int. J. Islam. Stud. Hum. Sci.*, vol. 4, no. 1, pp. 57–65, Jan. 2021, doi: 10.46722/hkmh.4.1.21c.

[12] C. O. Dávila, R. C. Lozoya, and S. Trabelsi, "Sociocultural influences for password definition: An AI-based study," *ICISSP 2021 - Proc. 7th Int. Conf. Inf. Syst. Secur. Priv.*, no. Icissp, pp. 542–549, 2021, doi: 10.5220/0010269305420549.

[13] Z. Xia, P. Yi, Y. Liu, B. Jiang, W. Wang, and T. Zhu, "GENPass: A Multi-Source Deep Learning Model for Password Guessing," *IEEE Trans. Multimed.*, vol. 22, no. 5, pp. 1323–1332, 2020, doi: 10.1109/TMM.2019.2940877.

[14] A. Bryk, "apriorit." https://www.apriorit.com/dev-blog/528-password-attacks-use-machine-learning.

[15] H. Xu *et al.*, "Adversarial Attacks and Defenses in Images, Graphs and Text: A Review," *Int. J. Autom. Comput.*, vol. 17, no. 2, pp. 151–178, 2020, doi: 10.1007/s11633-019-1211-x.

[16] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–11, 2015.

[17] R. Wiyatno and A. Xu, "Physical adversarial textures that fool visual object tracking," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2019-October, pp. 4821–4830, 2019, doi: 10.1109/ICCV.2019.00492.

[18] N. Akhtar and A. Mian, "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey," *IEEE Access*, vol. 6, no. August, pp. 14410–14430, 2018, doi: 10.1109/ACCESS.2018.2807385.

[19] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative Adversarial Networks: An Overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, 2018, doi: 10.1109/MSP.2017.2765202.

[20] I. Goodfellow *et al.*, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020, doi: 10.1145/3422622.

[21] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science (80-.).*, vol. 363, no. 6433, pp. 1287–1289, 2019, doi: 10.1126/science.aaw4399.

[22] E. Anthi, L. Williams, M. Rhode, P. Burnap, and A. Wedgbury, "Adversarial attacks on machine learning cybersecurity defences in Industrial Control Systems," *J. Inf. Secur. Appl.*, vol. 58, no. February, p. 102717, May 2021, doi: 10.1016/j.jisa.2020.102717.

[23] W. Brendel, J. Rauber, and M. Bethge, "Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models," *arXiv*, pp. 1–12, Dec. 2017, [Online]. Available: http://arxiv.org/abs/1712.04248.

[24] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical Black-Box Attacks against Machine Learning," in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, Apr. 2017, pp. 506–519, doi: 10.1145/3052973.3053009.

[25] S. A. Javadi, R. Cloete, J. Cobbe, M. S. A. Lee, and J. Singh, "Monitoring Misuse for Accountable 'Artificial Intelligence as a Service,'" in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Feb. 2020, pp. 300–306, doi: 10.1145/3375627.3375873.

[26] X. Liu, Y. Lin, H. Li, and J. Zhang, "Adversarial examples: Attacks on machine learning-based malware visualization detection methods," *arXiv*, no. September, 2018.

[27] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial Attacks and Defences: A Survey," *arXiv*, vol. x, no. x, 2018.

[28] S. Qiu, Q. Liu, S. Zhou, and C. Wu, "Review of artificial intelligence adversarial attack and defense technologies," *Appl. Sci.*, vol. 9, no. 5, 2019, doi: 10.3390/app9050909.

[29] H. Lauw, J. C. Shafer, R. Agrawal, and A. Ntoulas, "Homophily in the digital world: A livejournal case study," *IEEE Internet Comput.*, vol. 14, no. 2, pp. 15–23, 2010, doi: 10.1109/MIC.2010.25.

[30] G. S. O'Keeffe *et al.*, "Clinical report - The impact of social media on children, adolescents, and families," *Pediatrics*, vol. 127, no. 4, pp. 800–804, 2011, doi: 10.1542/peds.2011-0054.

[31] C. Salmivalli, "Bullying and the peer group: A review," *Aggress. Violent Behav.*, vol. 15, no. 2, pp. 112–120, 2010, doi: 10.1016/j.avb.2009.08.007.

[32] H. Sampasa-Kanyinga, P. Roumeliotis, and H. Xu, "Associations between cyberbullying and

school bullying victimization and suicidal ideation, plans and attempts among Canadian schoolchildren," *PLoS One*, vol. 9, no. 7, 2014, doi: 10.1371/journal.pone.0102145.

[33] E. Whittaker and R. M. Kowalski, "Cyberbullying Via Social Media," *J. Sch. Violence*, vol. 14, no. 1, pp. 11–29, 2015, doi: 10.1080/15388220.2014.949377.

[34] M. A. Al-Garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network," *Comput. Human Behav.*, vol. 63, pp. 433–443, 2016, doi: 10.1016/j.chb.2016.05.051.

[35] A. Kavanaugh *et al.*, "Social media use by government," p. 121, 2011, doi: 10.1145/2037556.2037574.

[36] J. M. Lampinen, J. Arnal, and J. L. Hicks, "The effectiveness of supermarket posters in helping to find missing children," *J. Interpers. Violence*, vol. 24, no. 3, pp. 406–423, 2009, doi: 10.1177/0886260508317184.

[37] J. Wolak, K. Mitchell, and D. Finkelhor, "Online Victimization of Youth: Five Years Later," *Juv. Justice*, p. 96, 2006, [Online]. Available: http://www.unh.edu/ccrc/pdf/CV138.pdf.

[38] C. Morris, "Identifying Online Sexual Predators by SVM Classification with Lexical and Behavioral Features 1," 2013.

[39] C. H. Ngejane, G. Mabuza-Hocquet, J. H. P. Eloff, and S. Lefophane, "Mitigating Online Sexual Grooming Cybercrime on Social Media Using Machine Learning: A Desktop Survey," *2018 Int. Conf. Adv. Big Data, Comput. Data Commun. Syst. icABCD 2018*, pp. 1–6, 2018, doi: 10.1109/ICABCD.2018.8465413.

[40] I. McGhee, J. Bayzick, A. Kontostathis, L. Edwards, A. McBride, and E. Jakubowski, "Learning to identify Internet sexual predation," *Int. J. Electron. Commer.*, vol. 15, no. 3, pp. 103–122, 2011, doi: 10.2753/JEC1086-4415150305.

[41] A. Kontostathis, L. Edwards, and A. Leatherman, "ChatCoder: Toward the tracking and categorization of internet predators," *Soc. Ind. Appl. Math. - 9th SIAM Int. Conf. Data Min. 2009, Proc. Appl. Math.*, vol. 3, pp. 1327–1334, 2009.

[42] D. T. L. Hui, C. W. Xin, and M. Khader, "Understanding the behavioral aspects of cyber sexual grooming," *Int. J. Police Sci. Manag.*, vol. 17, no. 1, pp. 40–49, 2015, doi: 10.1177/1461355714566782.

[43] B. Zhang, M. Anderljung, L. Kahn, N. Dreksler, M. C. Horowitz, and A. Dafoe, "Ethics and Governance of Artificial Intelligence Evidence from a Survey of Machine Learning Researchers," vol. 2019, no. 113, pp. 13–21, 2015.