DISAITEK

ML security

Updated 10 / 2019

Map of Threats over AI

Introduction

Why mastery over AI is important and what it will cost you if you are not aware of these topics.

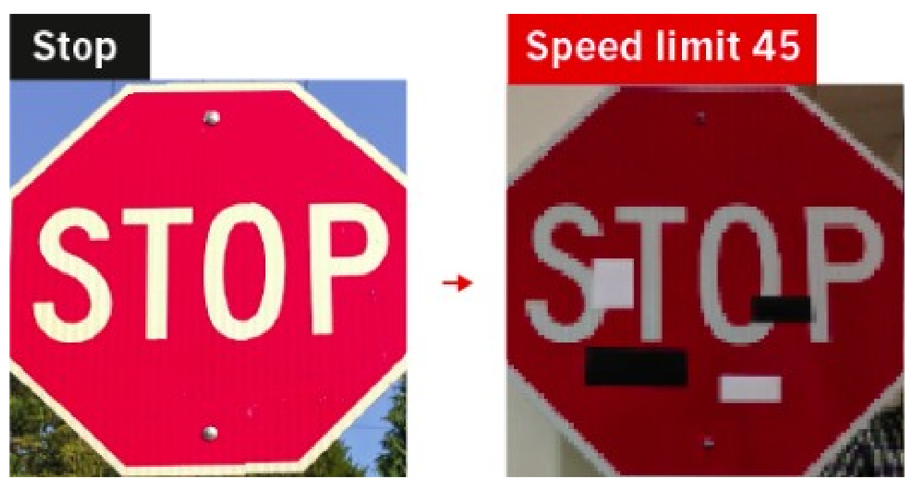
Machine Learning (*ML*), *Deep Learning* (*DL*) and more generally *Artificial Intelligence* (*AI*) is about to become a cog in most technological domains. The utility of *AI* is not to be demonstrated anymore. These technologies are currently *grey-box* algorithms from the point of view of the owner, because it is very hard to explain their internal functioning, foresee its limits and outlier cases. When put in production, theses systems may suffer, as other similar technologies, from the following issues:

- Intended or unintended misuse
- Data leaking, data governance issues
- Unintended & unfair biases (dealing with human related data for instance)
- Inability to deal with real unseen data

Moreover, unlike most technologies, AI often suffers from user mistrust, because it is not understood. So, in order for our society adopts this technology without friction, we need to demonstrate greater mastery, responsibility, liability and accountability over it. On the first hand, AI can be a tool for great prosperity but, on the other hand it can quickly become a root of an industrial and media disaster. In 2018, scandals linked to AI has never been so high in terms of number and magnitude. More than a direct cost, these scandals undermine the trust capital the public and potential clients have in *AI*, which slows down the progress, development and adoption of *AI*. It is a tremendous opportunity cost.

Security threats

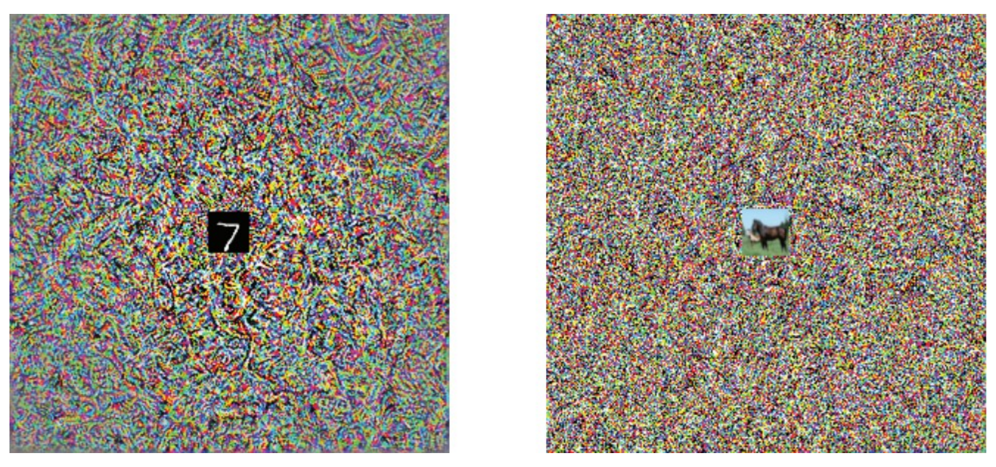
As any software with a public interface, *AI* is subject to the usual security threats like DoS and software vulnerability exploitation. Furthermore, it can be the target of threats at its algorithmic core. A carefully crafted input sample can be submitted to make the decision boundary of the trained AI change unnoticed. It is called an **adversarial sample**.



An adversarial sample fooling a Deep Neural Network trained to detect stop signs.

Deep Learning algorithms possess an interesting property from an attacker point of view: **transferability**. It is the phenomenon that two different deep learning models trained on the same problem exhibit similar decision boundaries. An attacker can use this property to train a deep learning algorithm on a similar problem, then find a working attack on its own model. This attack is likely to work on the target model without knowing its configuration or parameters.

A newly discovered threat is **adversarial reprograming**. This attack finds a single adversarial perturbation, that can be added to all test-time inputs to a machine learning model to cause the model to perform a task chosen by the adversary. It is GPU/CPU power stealing.



Adversarial program which cause ImageNet model to function as a MNIST classifier or a CIFAR-10 classifier.

Key takeaways

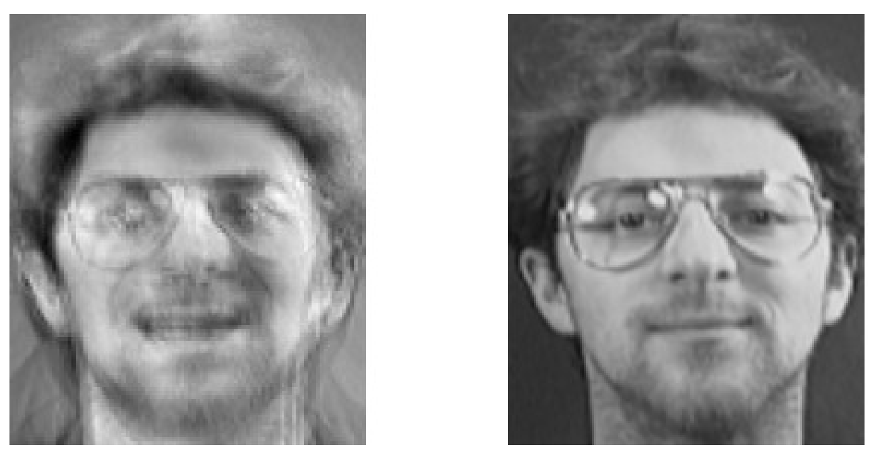
- Does your AI have a public interface?
  - It is subject to adversarial samples, adversarial reprograming, model stealing and sample reconstruction.
- Do you have a perfect control (access / content) of your training dataset?
  - Your AI is subject to data poisoning.
- Does your AI handle human-related data?
  - Your AI may base its decisions on unfair biases and can causes unintended discriminations.

In the two previous cases, we covered what can do an external attacker. In the following scenario, the attacker has access to the training samples and modify or add samples to partially control the machine learning model training. Either it is an inside job or the attacker gained access to the internal process. It is called **data poisoning**. The goal of the attacker is usually to shift the decision boundary of the model to its benefit (malware detector for instance).

Privacy threats

As Artificial Intelligence is tightly linked with Big Data, in the case these data are confidential or private, it can be a vector of attacks. Machine learning models « remember » partially their training samples. This capacity to remember training data can be exploited by testing if a sample has been part of the training set, this attack is called the **membership inference attack**. In the case where samples are linked to a person, such as medical or financial data, inferring whether samples come from the training dataset of the ML model constitutes a privacy threat. Meta-information could be leaked to the attacker.

Another attack caused by this phenomenon is **data extraction**. In this scenario, the attacker knows a part of a training sample and he is able to reconstruct the missing values by testing several of them on the trained model. The higher the confidence the model has in the result, higher the odds of the missing values being close to the original values. This attack is even possible when the attacker has to guess all values of the training sample :



An image recovered with a data extraction attack (left) vs the original (right). The attacker has only the person's name, can input sample and gathers the output of a facial recognition system.

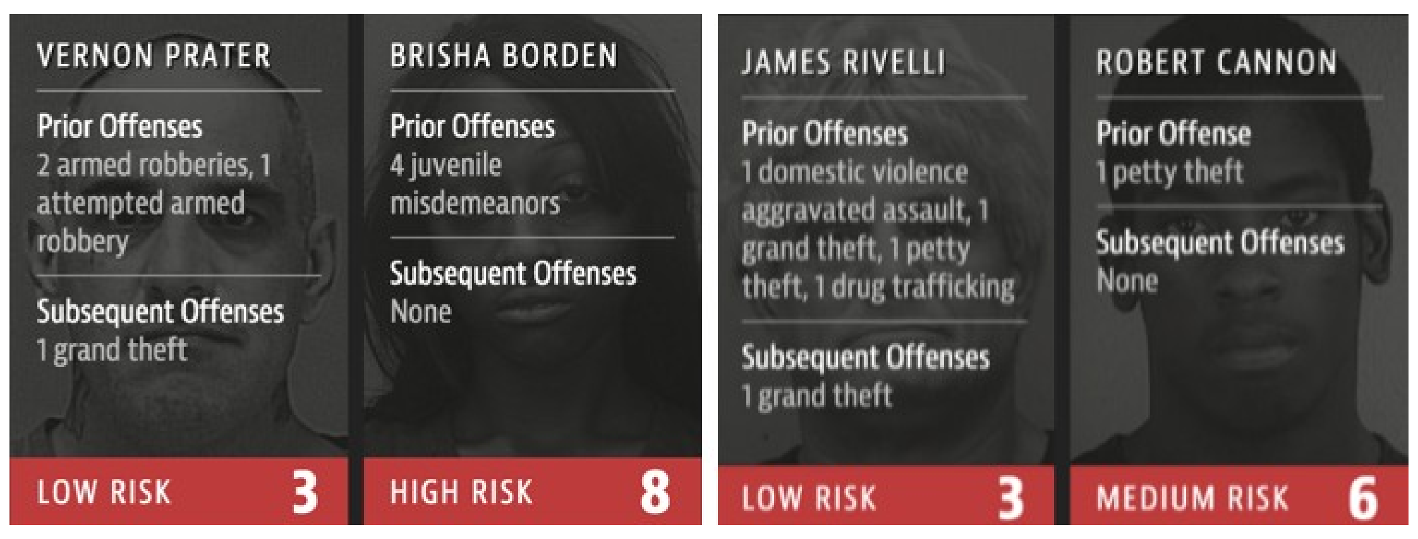
The last known attack is **model stealing**. In this scenario, the model is a priopprietary technology, and this model has a public interface. An attacker can forge or gather samples similar to the training samples and build a side model that mimics the predictions of the target model, effectively stealing it.

Unfair biases

The term bias in Machine Learning has different meanings in different contexts. Here, a bias is the importance that a feature has in the final model prediction. When the problem concerns people-related data, some features are unfit to participate to the prediction. These are the unfair biases. What is considered *fair* and *unfair* depends on the user's culture and country and is itself a subject to debate, but for sake of simplicity we are mainly refering here of the *gender* and *racial* biases.

Several scandals shown that machine learning reproduces or amplifies existing biases in the training dataset. For instance, Amazon uses a curriculum selection AI for recruiting. In 2018 a **Reuters article** exposes a serious problem of gender bias in this process. In fact, the model attached great importance to the words used more often by men and a negative importance to the words used more often by women. This is the result of data being unequally shared between men and women because there are the former greatly outnumber the latter at Amazon. The model has reproduced these inequalities in his predictions.

In 2016, the risk assessment AI software named COMPAS has been analyzed and a severe racial bias has been discovered. This tool was used to predict the risk of a criminal reoffending. Blackness ridiculously increased the predicted risk of the criminal.



COMPAS results

The general problem of unwanted biases is not only technological. It contains several topics that need to be discussed on a case-by-case basis. The list of unfair biases is local to the problem. For instance, in one case the age bias will be fair but in another case it will not be. There is also the question of how the fairness is measured that needs to be discussed; Different methods exist, which will not have the same relevance in all cases. This makes fairness a delicate subject on which there are both technological and political solutions.

Conclusion

Although Artificial Intelligence is a powerful technical solution, allowing incredible innovations, its implementation requires the mastery of a panel of subjects. Without this mastery, AI is a double-edged sword. The security, privacy and fairness of AI are subjects in construction that will be standard for the implementation of AI. In this context, we see initiatives emerging in this direction, such as the **Ethics guidelines for trustworthy AI** in the European Commission, the **OECD Principles on AI** or the adoption of **human-centered AI principles by the G20**. To be ahead of these topics is to be ahead of needs, standards and future threats.

About us

Disaitek was founded with a single mission: to use AI to bring knowledge and to bring knowledge over AI. We work on building trustworthy AI. Visit our website **https://www.mlsecurity.ai/** to see what we can do for your organization or contact us directly at **contact@disaitek.ai**.