Adversarial Machine Learning:

Approaches & defences

Prepared by:
Thomas Marcks von Würtemberg, Security Consultant
Matt Lewis, Research Director

Table of contents

1. Introduction

Most of us interact with Artificial Intelligence (AI) or Machine Learning (ML) on a daily basis without even knowing; from Google translate, to facial recognition software on our mobile phones and digital assistance in financial services or call centres. An IBM survey [1] of the automotive industry shows that 74 per cent of top executives expect to see self-driving cars, fundamentally operated through AI, by 2025.

AI is a growing market with ever increasing possibilities across all sectors, due to the vast amounts of 'big data' that most organisations now generate and accumulate. The utility of this data lies in its mining and use as training sets for AI/ML applications, becoming models on which to make predictions or decisions about future events that could be paramount to safety and/or security. As such, the integrity of data consumed by AI systems at both training and post-training phases is crucial. Similarly, the integrity of the AI algorithms themselves is paramount; depending on the algorithms or types of AI used, those algorithms may be susceptible to influence and change. From an adversarial perspective this might facilitate future compromise of safety or security measures, or simply serve to disrupt AI systems by conning them into making decisions that they might not otherwise make.

While many look forward to achieving goals with AI it is important to step back and reflect. AI has its critics; both Stephen Hawking and Elon Musk have voiced concerns about the dangers of AI. This has been primarily around concepts of algorithmic morality and where AI might eventually replace human, moralistic decision making, perhaps in ways that aren't conducive to peace, safety or preservation of planetary life.

In this paper we discuss 'Adversarial Machine Learning' and the potential impact of advances in this area of study. This is crucial in order to ultimately understand the defences or mitigations that can be put in place to minimise the risk of conning or tricking machines to manipulate their outputs and the decisions that they make. We look at where issues may arise, both during training and post-training, touching on concepts such as algorithmic morality and, where relevant, authority. From a broader AI assurance perspective we look at the issues around auditing AI systems and the data channels that they utilise, providing advice and guidance where possible on how the integrity of those data supply chains can be assessed and assured.

2. ML & adversarial ML

ML is a sub-field of AI and it focuses on the ability to train algorithms in order to improve their performance in solving a set tasks. One of the main areas in ML is classification, where algorithms are used to classify images or find patterns in large amounts of data. Note that the selection of training data is extremely important. An example [2] of this comes from the early days of AI when US army researchers wanted to classify Russian tanks from pictures. They applied ML to the problem and the test photos showed excellent results, however, when applied in real life the system did not work at all. The reason for this was that all of the test/training pictures of Russian tanks were taken on cloudy days, and all of the benign pictures were taken on sunny days, which made the model show all pictures of cloudy days as Russian tanks. In this paper we will take a slightly deeper look at these quirks and other types of problems that can occur with ML.

Another feature that makes ML such a powerful tool is that it is possible to retrain algorithms to adapt to different environments or changes in datasets, which is a useful feature in the modern and ever-changing world. An example of this is intrusion detection software, where the state of normal might differ from time to time or network to network. However, with ML a product can be trained on the specific environment, harmonising detection and minimising false positives. A more in-depth discussion on the subject of ML and AI in cyber security is available in the NCC Group whitepaper “Rise of the machines: Machine Learning & its cyber security applications” [3].

Adversarial ML looks at the possibility of an attack on, or compromise of, systems that employ ML. This is usually done through manipulation of the data inputs to systems based on ML, either during the training/learning phase, and/or operational phase. From an adversary’s perspective, they may be keen to identify possible inputs that would be falsely classified (perhaps to evade a detection system), or they might be interested in forcing the ML-based system to adapt in ways so that it eventually and consistently misclassifies data or learns unintended or bad behaviours that are to the adversarial advantage.

3. Current attacks against ML

In the last eight years a new area of research has emerged, looking at how ML algorithms can be attacked and conned into incorrect classification or decisions. Since ML works on the basis of statistics, we can use this to push certain mathematical values high, which brings up averages and leads to incorrect results within the ML algorithm. Current research on this type of attack is mostly focused around connected/smart cars and their ability to read road signs and react to their environment. One study showed that it was possible to add pixels to digital images in order to push the threshold of a classifier in the wrong direction [4]. Another study showed the possibility of such attacks surviving even after the image was printed [5] onto paper and then classified.

There has been some debate about the practicality of such adversarial attacks in the physical world; the possibilities are much more limited when it comes to adding adversarial data and even at what angle the camera sees the object it tries to classify. These physical world attacks have, however, been demonstrated in a paper which demonstrated that it is remarkably easy and simple to trick classifiers, even by something as simple as graffiti or stickers on a stop sign [6].

Another issue that arises from ML and AI is the overconfidence in their ability and to blindly trust them to take correct and informed decisions. Today, ML is based on statistical models and it has been well-proven that it is possible to trick ML systems with tainted data. Since the algorithms are self-taught we have little to no insight into how they work, or what makes them classify things correctly. They are “oracles” or “black boxes” to us as humans, which given an input produces an answer. This is one of the big drawbacks of ML, as since we do not know all of their inner workings, it is hard to audit them using conventional testing or assessment methods. For ML applications in security we may struggle to understand how to fully guard against, or detect, adversarial malformed inputs that seek to manipulate the underlying algorithms and the decisions that they make.

ML is frequently used for tasks like speech recognition, but even here attacks have been demonstrated. Most notable is a study from 2016 [7] which showed that it was possible to covertly send commands to phones via inaudible ranges. The human hearing range for sound is from 20 Hz to 20 kHz. Computers, however, have the ability to pick up a broader spectrum of frequency, dependent on the microphone in use. This property was successfully exploited by a group of researchers who showed it was possible to send commands to a phone’s voice recognition system, which the phone’s internal algorithm classified as commands. This was despite the fact that they were outside or just on the edge of the human hearing spectrum.

3.1 Algorithmic morality & authority

As ML and data-driven models become more commonplace in products there are areas that need to be addressed with regards to ethics in algorithms (algorithmic morality). The old trolley problem [8] can be applied to self-driving cars, as they will face this exact ethical dilemma, but there are many more complex problems around ethics and algorithms. We like to think of computers as cold logical machines, but when introducing human programming and data selection, one can almost never avoid human bias occurring. One area where this is important, for both ethical and legal reasons, is algorithmic hiring (in recruitment). Where questions have been raised [9], but not yet answered, is that what happens if an algorithm that chooses employee candidates begins to pick up undesirable patterns. This could include not selecting women who are more likely to take maternity leave, or even deselecting candidates that have a higher likelihood of suicide? These questions are incredibly valid as ML is extremely effective at picking up patterns, even ones that we as humans don't see. This has been demonstrated in studies around fall detection in elderly people [10]. This area has not seen the same advances in research as the problems around self-driving cars, and more research is needed.

Algorithmic authority concerns our behaviour as humans to trust and follow the outputs produced by ML-based systems, as if those decisions were more authoritative than any that could be made by humans operating within the same problem space. That is we may welcome the authority derived from the superior processing power and huge datasets used to train ML-based systems. As one example, the airline industry has introduced computer-based automation in recent years. Modern aeroplanes can fly themselves, but they are constantly supervised by pilots that are ready to take over if something fails with the AI. However, one of the leading causes of modern aviation accidents is pilots being overconfident in the automation of their aircraft. The same concerns and questions arise in cyber security - are we putting too much trust and too little human supervision in areas of automation such as malware or breach detection for example?

Note that computational issues around algorithmic morality and authority might not be anything to do with adversaries, but rather by design and evolution the algorithms may learn behaviours that we humans would consider immoral. However, due to our inherent trust in their operation, we might simply bow to their algorithmic authority in the belief that their decisions are superior to those that we mere mortals could formulate. This speaks more to a risk and limitation of human nature as opposed to a technical or machine-based risk, but human supervision and authority override may be key to ensuring safe and moral ML-based system operation.

A number of useful resources that cover technical issues aligned to the safe development and implementation of ML/AI can be found at OpenAI [11], a non-profit AI research company, discovering and enacting the path to safe artificial intelligence.

3.2 ML subversion & accessibility of techniques

The reality of subverting ML-based systems rests with us, and therefore this helps us to maintain perspective around any associated hype. While ML-based applications can work well to their assigned task(s) they are, by design, susceptible to subversion, and there is growing, emergent research from both academia and industry [12] in this domain. This includes work on Generative Adversarial Networks (GANs), which involves the use of neural networks in actual adversarial training while taking part in a zero-sum game [13].

There are a growing number of online resources available to support adversarial ML tasks. While ML concepts can be difficult to comprehend due to the requirement for a strong background in mathematics and statistics, emerging applications are helping to abstract away from these details. This makes attacks against ML systems ever more likely and more accessible to the less mathematically/ML-inclined adversaries. Examples include Deep-Pwning, termed the “metasploit for machine learning”. This ever evolving framework, released at DEFCON 24, is presented as a “lightweight framework for experimenting with ML models with the goal of evaluating their robustness against a motivated adversary.” [14]

OpenAI has also documented successful adversarial samples [15], while a number of resources exist such as curated resources on adversarial ML [16].

4. Evaluating ML

Systems employing ML need to be assessed and evaluated, particularly to validate vendor claims around performance. The core problem here, however, is the black box nature of ML and AI systems. While static analysis of algorithms can be performed, it is the dynamic nature of those algorithms that provides the utility of the system, and therefore it is the dynamic aspect which must be assessed. However, this may commonly comprise hidden layers, such as within neural networks. A suggested approach to ensure oversight of ML/AI algorithms (algorithmic transparency) is algorithmic auditing. An MIT Technology Review article noted 'auditability' as one of the five principles for accountable algorithms [17]. The suggestion is that auditability should be 'baked in' to algorithms in their development stage in order to enable third parties to check, monitor, review and critique their behaviour. This also satisfies peripheral regulatory requirements, such as data protection, where perhaps the data processed by the ML solution is personal in nature [18].

The best method available for evaluating ML is therefore to exploit the black box approach by understanding what the task of the application is, firing through various samples to assess the robustness of the classification/decision making as an output. This is in terms of both false positives and false negatives. For these tests we can employ the tools and techniques described around adversarial ML; that is, use modified or mutated inputs to assess the system's ability to properly classify (or misclassify). Work here should also include assessment of the original training data and how current it is/was, since the age of training data could be an initial indication of the likely system performance. A recent example here is of false positive reporting, seen through the misclassification of a benign "Hello World" binary as malicious by various ML-based malware engines [19]. From reviewing various vendor responses, a common theme appears to relate to how the engines had been trained. This shows the importance of training robustly, and/or periodic re-training so that classifiers can remain current and abreast of changes in possible input samples.

Much work has already been done around metrics for assessing ML algorithm performance [20]. Over time we will surely see frameworks and test harnesses that will allow for automated, repeatable and consistent testing of ML-based systems, leveraging these different types of classification metrics.

5. Defence & mitigation

The selection of the appropriate and diverse training data is a key issue relating to ML. Many current mitigations against adversarial attacks are based on adding random adversarial data to the training set in order to encourage the model to classify correctly, even with the extra data. However, as we add more data we introduce issues around not knowing exactly what this complex model is triggering, as with the example of the US army tank classification [2].

The most recent research in defence [21] is around detecting and rejecting the dangerous data before it reaches the classifier. This is done by wrapping the classifier with an algorithm that handles data sanitation. A positive side of this approach is that it works regardless of what classifier algorithm is used, making it easy to adapt to current use cases. A major drawback here is understanding what constitutes 'malicious inputs'. If inputs are within permissible data value ranges then sanitising this data may prove difficult, if not impossible.

Potential defences are also emerging from academic research – this includes actual adversarial training on a large scale [22] and techniques for detecting concept drift in ML-based systems [23].

6. Conclusion

ML is a powerful tool, yet despite its widespread use there are many unknowns. This is not only in terms of how it actually works, but also what makes the system trigger or classify on a certain sample for example. Most ML products in use today are black box appliances that are placed onto networks and configured to consume data, process it, and output decisions, without us humans having much knowledge on what those appliances are actually doing. Apart from the risk of being overconfident and trusting these devices to make the correct decisions, there might in some cases also be legal issues in trusting the outcome of an algorithm. New regulations and research into the area of algorithmic authority and the legality of decisions made by machines are emerging and will hopefully bring future clarity in this domain.

ML-based systems can be conned and there is constant, emerging research and accessible toolsets that in time will likely aid adversaries in their attempts at ML subversion, while on the periphery, we are also seeing counter-research around detection and blocking of adversarial attacks. The core issue of data taint during training and/or operation will always exist and pose some level of risk to the underlying ML-based system and the decisions it makes, which links to a major issue around data supply chain. Being able to assure the origin and integrity of all data passed to ML-based systems may be impossible, particularly in large-scale enterprise systems that consume large volumes of data, such as network log information for example. This means that adversaries may have a myriad of vectors available to them to attempt manipulation of data that might ultimately affect ML-based system operation.

Many ML product vendors are, understandably, protective of their Intellectual Property Rights (IPR), which could cause issues around technology and security audits. Various regulators in different sectors may not be satisfied with a response of “it’s a black box” when attempting to audit systems, particularly if those systems instigate key decisions such as financial transactions or security operations. This is where concepts such as algorithmic transparency will become ever more important, requiring vendors to build in the necessary assurances and auditing capabilities to their ML-based products.

The security industry has been quick to adapt the use of ML as it is well suited to the problem of classifying data from large data sets. Today, there are a number of products within the areas of spam filtering, malware detection and network intrusion detection which use ML at their core. There are many claims made by ML product vendors about their effectiveness in detecting threats and while these may be true, they are often lacking in any independent third party validation. Therefore the claims should not be taken as gospel without adequate and independent validation.

As the cyber security industry has witnessed and learned with Intrusion Detection Systems (IDS), when detection signatures become outdated and are not maintained with signatures of the latest, emerging threats, the detection mechanism itself becomes less effective. The same issues may apply to ML-based systems whereupon their models/classifiers become outdated and ineffective. More research is needed to understand the appropriate frequencies of model updates for different ML-based systems (e.g. every week, month or quarter). Additionally, we need to improve our understanding of how best to perform dynamic model updating and how much supervision vs. authority should be placed in any such dynamic, algorithmic update.

There is also an emerging issue around abstraction with regards to ML. There are an ever-growing number of ML/AI frameworks (including open source) becoming available to software developers that abstract away the data science and algorithmic details within their application programming interfaces (APIs).This means developers don't necessarily need to understand the underlying mathematics and may therefore be making poor choices in terms of algorithm, labels and attributes used for classification, or the types of dimension reduction employed etc.

Finally, we conclude with ten questions that we recommend those responsible for implementing ML-based systems ask in relation to the problem they are aiming to solve, the target operating environment and the chosen ML system vendor(s). Answers to these questions should help to address concerns, focus priorities on risk reduction activities and to understand any gaps in legal or regulatory compliance. This list of questions is by no means exhaustive and, instead, should serve as a good starting point, particularly for those new to AI/ML-based concepts and systems.

In the interest of humanity, we also suggest that these questions are asked of, and answered by, humans. That is, we do not recommend posing these questions to the underlying machines themselves, or creating machines to learn and answer these questions.

1. What datasets were used for training the system and how appropriate/current were these for the target operating environment?

2. How often does the system retrain, or how often should it be retrained?

3. What level of testing/evaluation has been done on false positive and false negative rates?

4. What (if any) adversarial testing has been performed against the system?

5. What (if any) defence mechanisms exist, such as adversarial training/GANs, concept drift detection, etc.?

6. What level of Algorithmic Authority will the system present, or what level of trust will be placed in the system's decision making?

7. What level of human supervision and authority override will be required during system operation?

8. Are there any Algorithmic Morality issues presented by the system?

9. What level of Algorithmic Transparency and Auditing exists?

10. What assurances exist in the data supply chain?

In October 2017, the University of Berkeley presented their collective view of systems challenges for AI [24] which echoed some of the potential issues raised in this paper. Also in October 2017, in the UK an independent review [25] published on behalf of the AI sector presented the opportunities for the UK to become a clear world leader in the development of AI, by helping to boost productivity, advance health care, improve services for customers and unlock £630bn for the UK economy. It is hoped that the topics in this paper have served as food for careful thought around the adversarial aspects of AI so that AI as a sector can be cognisant of what needs to be done to viably realise such world-leading, world-impacting prospects.

References

[1] http://www-935.ibm.com/services/multimedia/GBE03640USEN.pdf

[2] http://intelligence.org/files/AIPosNegFactor.pdf

[3] https://www.nccgroup.trust/uk/our-research/rise-of-the-machines-machine-learning-and-its-cyber-security-applications/

[4] https://arxiv.org/abs/1602.02697

[5] https://arxiv.org/abs/1607.02533

[6] https://arxiv.org/pdf/1707.08945.pdf

[7] https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/carlini

[8] https://en.wikipedia.org/wiki/Trolley_problem

[9] https://www.ted.com/talks/zeynep_tufekci_machine_intelligence_makes_human_morals_more_important

[10] www.mdpi.com/1424-8220/14/10/19806/pdf

[11] https://openai.com/about/

[12] https://www.slideshare.net/mobile/RamShankaraSivaKumar/subverting-machine-learning-detections-for-fun-and-profit

[13] https://github.com/openai/InfoGAN and https://arxiv.org/abs/1606.03657

[14] https://github.com/cchio/deep-pwning and https://media.defcon.org/DEF%20CON%2024/DEF%20CON%2024%20presentations/DEFCON-24-Clarence-Chio-Machine-Duping-101.pdf

[15] https://blog.openai.com/robust-adversarial-inputs/

[16] https://github.com/yenchenlin/awesome-adversarial-machine-learning

[17] https://www.technologyreview.com/s/602933/how-to-hold-algorithms-accountable/

[18] https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf

[19] http://www.csoonline.com/article/3216765/security/heres-why-the-scanners-on-virustotal-flagged-hello-world-as-harmful.html?page=2

[20] https://machinelearningmastery.com/metrics-evaluate-machine-learning-algorithms-python/

[21] https://arxiv.org/abs/1705.09064

[22] https://arxiv.org/abs/1611.01236

[23] https://s2lab.isg.rhul.ac.uk/papers/files/usenixsec2017.pdf

[23] https://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-159.pdf

[24] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/652097/Growing_the_artificial_intelligence_industry_in_the_UK.pdf

[1] http://www-935.ibm.com/services/multimedia/GBE03640USEN.pdf

[2] http://intelligence.org/files/AIPosNegFactor.pdf

[3] https://www.nccgroup.trust/uk/our-research/rise-of-the-machines-machine-learning-and-its-cyber-security-applications/

[4] https://arxiv.org/abs/1602.02697

[5] https://arxiv.org/abs/1607.02533

[6] https://arxiv.org/pdf/1707.08945.pdf

[7] https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/carlini

[8] https://en.wikipedia.org/wiki/Trolley_problem

[9] https://www.ted.com/talks/zeynep_tufekci_machine_intelligence_makes_human_morals_more_important

[10] www.mdpi.com/1424-8220/14/10/19806/pdf

[11] https://openai.com/about/

[12] https://www.slideshare.net/mobile/RamShankaraSivaKumar/subverting-machine-learning-detections-for-fun-and-profit

[13] https://github.com/openai/InfoGAN and https://arxiv.org/abs/1606.03657

[14] https://github.com/cchio/deep-pwning and https://media.defcon.org/DEF%20CON%2024/DEF%20CON%2024%20presentations/DEFCON-24-Clarence-Chio-Machine-Duping-101.pdf

[15] https://blog.openai.com/robust-adversarial-inputs/

[16] https://github.com/yenchenlin/awesome-adversarial-machine-learning

[17] https://www.technologyreview.com/s/602933/how-to-hold-algorithms-accountable/

[18] https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf

[19] http://www.csoonline.com/article/3216765/security/heres-why-the-scanners-on-virustotal-flagged-hello-world-as-harmful.html?page=2

[20] https://machinelearningmastery.com/metrics-evaluate-machine-learning-algorithms-python/

[21] https://arxiv.org/abs/1705.09064

[22] https://arxiv.org/abs/1611.01236

[23] https://s2lab.isg.rhul.ac.uk/papers/files/usenixsec2017.pdf

[24] https://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-159.pdf

[25] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/652097/Growing_the_artificial_intelligence_industry_in_the_UK.pdf