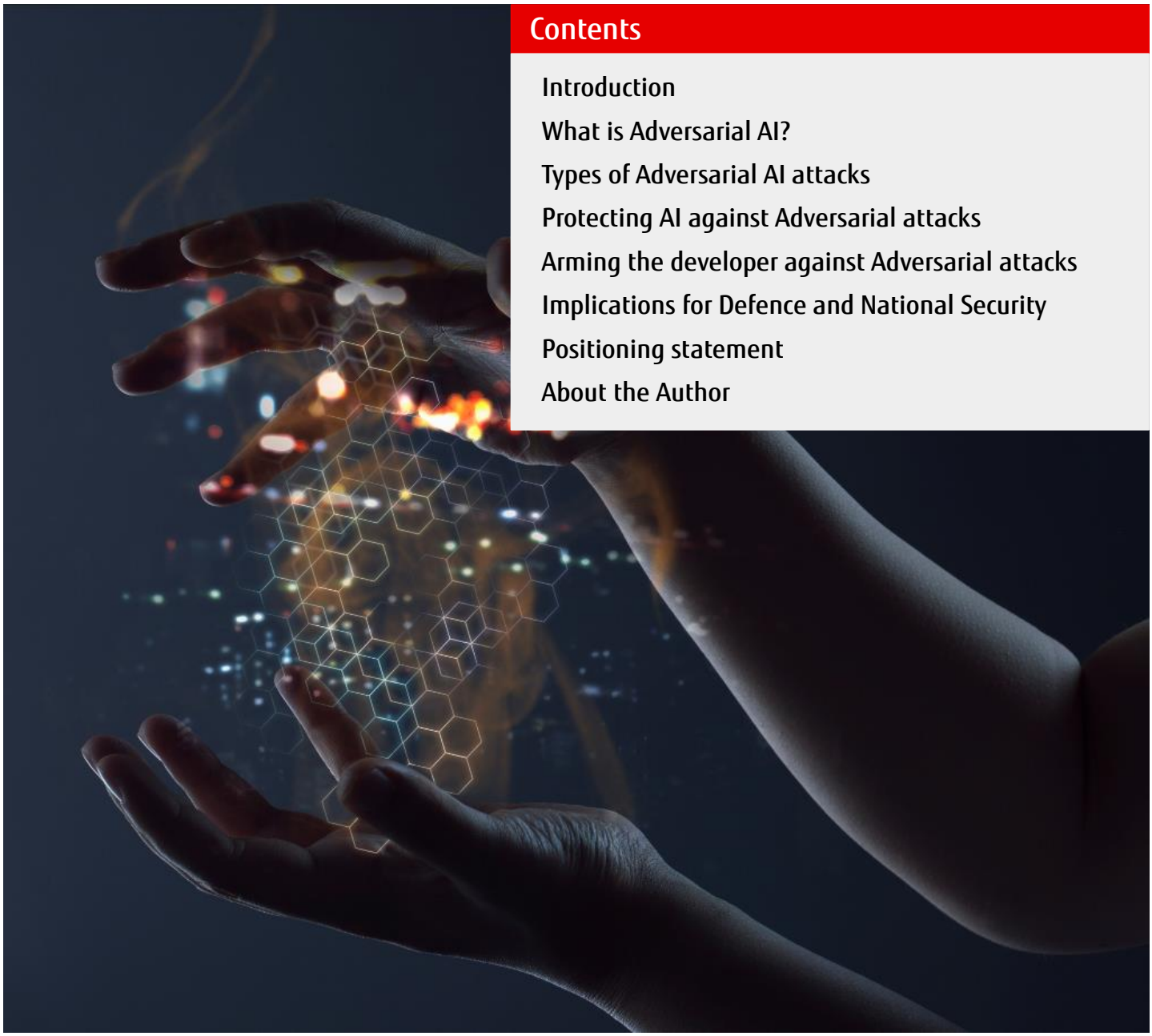
FUJITSU

Adversarial AI

Fooling the Algorithm in the Age of Autonomy

As Artificial Intelligence, or AI, becomes further embedded into the economic and social fabric of our day-to-day lives, maintaining the integrity of these systems and the data they use is paramount.



Contents

Introduction

Growing pervasiveness has given unscrupulous attackers the opportunities to exploit any vulnerability found within machine learning models and the data used to train them, giving rise to 'Adversarial AI'. The potential impact that Adversarial AI can have on our society and the harmful implications it will create to our security, trust and general wellbeing will only become more apparent with the continued pace in adoption of autonomous systems.

So, what can we do to mitigate and plan against these risks?

What is Adversarial AI?

The idea of an Adversarial AI attack is fundamentally very simple. An attacker can look to generate and introduce small changes to a dataset that, although imperceptible to the human eye, can cause major changes to the output of an AI system. Adversarial AI causes machine learning models to misinterpret the data inputs that feed it. As a result, it makes it behave in a way that's favourable to the attacker.

To produce unexpected behaviour, attackers create 'adversarial examples'. These often resemble normal inputs, but instead are meticulously optimised to break the model's performance. Attackers typically create these Adversarial examples by developing models that can repeatedly make minute changes to the data inputs of an AI system. These are often known as 'poisoning attacks', where the machine learning model itself then becomes compromised[1][2].

A good example where poisoning attacks can have significant implications is in image classification systems. An Adversarial attacker can introduce random noise into input image datasets that completely alters the results of a trained classifier[3]. While this might sound like a 'fun' research exercise, imagine the damage it can cause in scenarios such as self-driving vehicles. Attackers could target autonomous vehicles by placing stickers or using paint to create an Adversarial stop sign that the vehicle would interpret as another type of sign[4].

Clearly, Adversarial AI has the potential to become a major security threat. If an adversary can determine a particular behaviour in a model that's unknown to system developers, they can look to exploit that behaviour in order to create intentional consequences.

Types of Adversarial AI attacks

Most of the innovation in AI today has come about through the application of deep neural networks. The majority of all deep neural networks are trained to optimise their behaviour in relationship to a specific task, such as language translation or image classification. During training, this desired behaviour is usually formulated as an optimisation problem which minimises a known loss function by measuring the deviations from desired behaviour.

Adversarial attacks create input examples that seek to maximise this loss value and consequently maximise the deviations from desired behaviour[5]. However, creating the correct Adversarial examples requires prior knowledge of the inner workings of the deep neural network model. Fundamentally, attackers approach this problem using two types of strategy:

- **White-box attack:** the strong assumption here is that the adversary has full knowledge of the inner workings of the deep neural network and can utilise this knowledge to design 'adversarial examples'.
- **Black-box attack:** the adversary has a limited knowledge of the architecture of the deep neural network and can only estimate the behaviour of the model and devise Adversarial examples based on this estimation.

Whichever type of strategy is used, there are currently five known Adversarial techniques that can be leveraged and repeated against deep learning models:

1. **Evasion:** Attacks that modify the input to influence the model e.g. adding modifications to images in order to influence classification. This technique can be used to evade a model to correctly classify situations in a downstream task.
2. **Model Poisoning:** Adversaries can train machine learning models that are performant, but contain backdoors that produce inference errors when presented with inputs containing a trigger defined by the adversary. This backdoor model can be exploited at inference time with an evasion Attack.
3. **Training Data:** Attacks that modify training data add another backdoor e.g. imperceptible patterns in training data create backdoors that can control model outputs.
4. **Extraction:** Attacks that steal a proprietary model e.g. attacks can launch queries against a model regularly in order to extract valuable information to reveal its properties.
5. **Inference:** Attacks that earn information on private data e.g. an attack.

[1] https://www.aimagazine.com/data-and-analytics/data-poisoning-new-front-ai-cyber-war
[2] https://bdtechtalks.com/2019/04/29/ai-audio-adversarial-examples/
[3] https://venturebeat.com/2020/02/24/googles-ai-detects-adversarial-attacks-against-image-classifiers/
[4] https://towardsdatascience.com/your-car-may-not-know-when-to-stop-adversarial-attacks-against-autonomous-vehicles-a16df91511f4
[5] https://portswigger.net/daily-swig/adversarial-attacks-against-machine-learning-systems-everything-you-need-to-know

Protecting AI against Adversarial attacks

Protecting AI against such Adversarial attacks is challenging and, similar to cyber security, suffers from the limitation of assuming prior knowledge of attacks, which is not ideal for real-world scenarios. Adversarial techniques are constantly evolving, and bad actors regularly develop new attack methods, causing AI systems to face attacks that haven't been evaluated during their training phase. What makes Adversarial attacks different from cyber threats, however, is their unknown nature and the possible countermeasures.

For most security vulnerabilities, the boundaries are very clear. Once a bug is found, security analysts can precisely document the conditions under which it occurs and find the part of the source code that is causing it. Appropriate patches can be applied accordingly. Understandably, given the statistical nature of Adversarial attacks, it's difficult to address them with the same methods used against code-based vulnerabilities as you can't point to the exact line of code that is causing the vulnerability, since it spreads across the thousands and millions of parameters that constitute the AI model. As a result, evaluating the robustness of an AI system against unforeseen Adversarial attacks has become an increasingly important research topic.

Until now, Adversarial defences have involved an element of statistical adjustment or general changes to the architecture of the machine learning model. One emerging approach being explored is Adversarial training, where researchers probe a model to produce Adversarial examples and then retrain the model on those examples and their correct labels [6]. Adversarial training then readjusts all the parameters of the model to make it robust against the types of Adversarial examples it has been trained on.

The biggest challenge that arises from this is to produce enough varied and wide-ranging examples in order to combat the uncertain nature of Adversarial attacks. Utilising simulation/synthetic data to train models with numerous and diverse Adversarial examples mimicking a wide range of distortion sizes is also seen as an approach for allowing scalable evaluation of AI systems [7]. Such frameworks can provide a necessary route towards measuring model robustness against unforeseen Adversarial attacks. It also supports building a catalogue of Adversarial attack-based strategies to strengthen future model development.

Arming the developer against Adversarial attacks

While it is recognised that the tools and procedures for defending AI systems against Adversarial attacks are still in their preliminary stages, there a number of things that can be done now from an AI developer's perspective. Recently, the Adversarial ML Threat Matrix[8], published by researchers at Microsoft, IBM, Nvidia, MITRE, and other security and AI companies, provides security developers with a framework to find weak spots and potential Adversarial vulnerabilities in software ecosystems including machine learning components. The two key recommendations made were:

- **Evaluating the robustness of machine learning models** against Adversarial attacks as an integral step in model development and continuous integration processes (testing). This will encourage the rapid crafting and analysis of attack and defence methods for machine learning models. It will also generate metrics about the robustness of any trained model before deployment.
- Adversarial defence tools that will be developed in the future need to **be backed by the right policies** to make sure machine learning models are safe. Software platforms - such as GitHub, which developers commonly use to open source their work - must establish procedures and integrate these tools into the vetting process of applications that will use these machine learning models.

[6] https://portswigger.net/daily-swig/adversarial-attacks-against-machine-learning-systems-everything-you-need-to-know
[7] https://blog.global.fujitsu.com/fgb/2020-07-03/tackling-the-ai-data-challenge-could-synthetic-data-be-the-answer/
[8] https://github.com/mitre/advmlthreatmatrix

Implications for Defence and National Security

Adversarial attacks also pose a tangible threat to the stability and safety of AI and robotic technologies, which are increasingly going to be incorporated into defence and national security systems[9]. Defence has to ensure that the testing of AI systems against Adversarial attacks is a key requirement that becomes embedded within the lifecycle and maintenance of mission-critical applications.

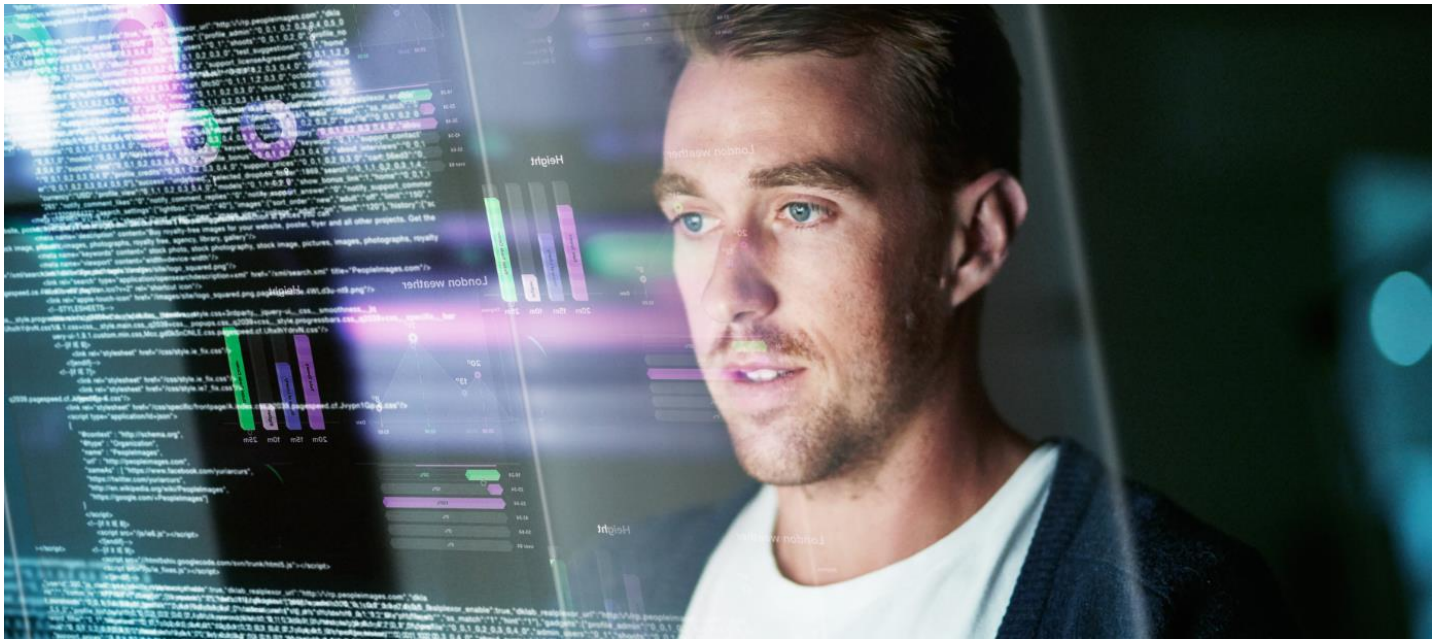
The challenge for Defence - like many other commercial entities - is knowing the exact conditions for such attacks. These are typically quite unintuitive for humans and notoriously difficult to predict when and where the attacks could occur.

Increasingly hostile actors have been known to employ reconnaissance-based techniques to understand attack strategies through leveraging publicly available information, or Open Source Intelligence (OSINT), about an organisation[10]. This information could help to identify where or how machine learning is being used in a system, and help tailor an attack to make it more effective. Common sources of information typically include technical publications, blog posts, press releases, software and public data repositories, as well as social media posts.

While it may be possible to estimate the likelihood of an Adversarial attack, the challenges of knowing the exact response the AI system will take is also difficult to predict. This has the potential to lead to outcomes where less safe military engagements and interactions are a result, and trust is compromised. Catering for these types of scenarios - and possibly more - will have to be verified and used where applicable to guide suitable response mechanisms.

Testing for all potential scenarios may not be possible, especially where AI technologies are being increasingly used to handle the enormous volumes of data generated for Intelligence, Surveillance and Reconnaissance (ISR) purposes. AI technologies for ISR will play a significant role in the creation and maintenance of situational awareness for human decision-makers at the edge. In such situations, the destabilising risks of Adversarial attacks will again be of some concern.

Defence should look to take a 'Systems of Systems' design approach towards understanding the impact of compromised information from the edge on the overall stability of the decision support system, and to ensure that parts of the critical chain have suitable checks and balances[11]. A risk analysis of the combined system (risk aggregation) may indicate and reveal a different risk impact and risk likelihood. Where this cannot be assessed in real operational environments prior to use, simulation offers a different route for providing the evidence required to allow continued operation with the necessary system assurance guarantees [12].



[9] https://www.gov.uk/government/publications/human-machine-teaming-jcn-118
[10] https://www.lexology.com/library/detail.aspx?g=cc79aa80-0164-42ca-a21a-22226ef0e7e3
[11] https://www.isaca.org/resources/news-and-trends/newsletters/atisaca/2019/volume-17/systems-thinking-in-risk-management
[12] https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00691/full

A 'systems of systems' approach also enables Defence to further assess and potentially limit the impact of compromised information by evaluating the role of humans, in terms of, where they should be best placed within the decision-making loop to ascertain information reliability. Within the context of human machine teaming[13], humans could potentially be trained to monitor such attacks and to assist the guidance of AI systems to more appropriate behaviours against known safety bounds. This is to ensure, if and when required, that any future operator be best placed with the correct situational awareness in order to be able to take full control of the AI system, both safely and effectively.

Adversarial attacks will also impact and exacerbate the co-ordination of decision-making challenges frequently associated with multinational military operations carried out by allies and security partners. Policymakers and experts in the United States and other countries have urged international co-operation on the development and use of AI, including the defence against Adversarial AI from hostile actors.

The US Defense Department, Joint Artificial Intelligence Center (JAIC), is partnering with like-minded nations to solve global security challenges and technological innovations. Initiatives like this illustrates the thinking and need to work more closely in areas such as AI to counter hostile actions[14].

Positioning Statement

At Fujitsu, we are acutely aware of the potential risks posed by Adversarial AI. As such, we are closely engaged with industry, academia and regulators as they continue to investigate and develop good practise, measures and guidelines to ensure that appropriate defences to Adversarial attacks are built into AI solutions from across a wide range of industry applications.

About the Author

Dr. Darminder Ghataoura has over 15 years' experience in the design and development of AI systems and services across the UK Public and Defence sectors as well as UK and international commercial businesses. Darminder currently leads Fujitsu's offerings and capabilities in AI and Data Science within the Defence and National Security space, acting as Technical Design Authority with responsibility for shaping proposals and development of integrated AI solutions. He also manages the strategic technical AI relationships with partners and UK government.

Darminder holds an Engineering Doctorate (EngD) in Autonomous Military Sensor Networks for Surveillance Applications, from University College London (UCL).

[13] https://www.gov.uk/government/publications/human-machine-teaming-jcn-118
[14] https://www.lexology.com/library/detail.aspx?g=cc79aa80-0164-42ca-a21a-22226ef0e7e3

FUJITSU

Address: Jays Close, Basingstoke, Hampshire, RG22 4BY
Tel: +44 (0) 203 949 3983
Email: helen.tilsley@uk.fujitsu.com
Web: uk.fujitsu.com