EXPOSING THE HYPE: THE ROLE OF AI IN CYBER SECURITY

**By Richard Walters**
CTO, Censornet

ABOUT THE AUTHOR

Richard Walters has 30 years of experience in the IT and security industry as a co-founder, innovator and security product leader, including over 15 years in C-level positions focused on information security.

Richard has in-depth knowledge of operating system and database security, intrusion detection systems, email and web security, identity and access management, and cloud and mobile security. He also spent over 10 years consulting with FTSE100 companies on all aspects of security and providing corporate intelligence on evolving threats from criminal and state actors.


Richard Walters, CTO, Censornet
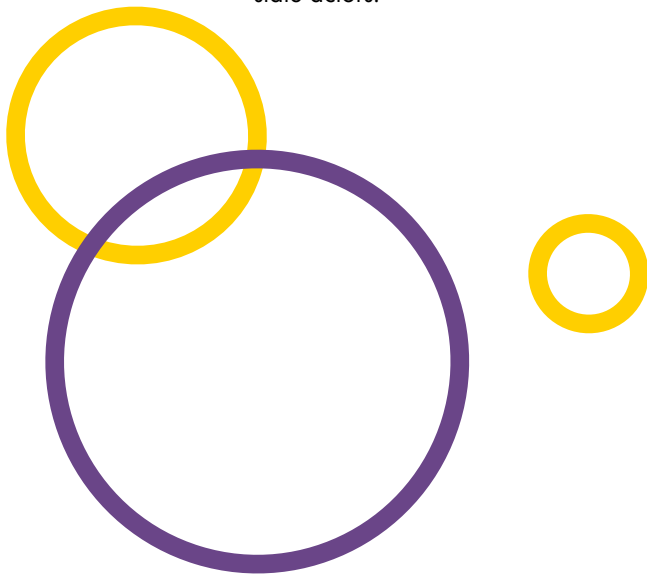
ABOUT KONCISE

We wanted to team up with our vendor Censornet and produce this white paper on such a significant topic. AI is being spoken about more and more in the cyber security space, so we wanted to join the conversation.

As a cyber security re-seller, we are less laser-focused on technical delivery and the hard sell and focused more on a double-handed approach.

We talk to you about ideas – yours and ours – as to how we can work together in the long-term to ensure we spot opportunities others may miss and thinking you may not have considered. The approach we take to IT and how it impacts your business is grounded in original thinking, world-class, award-winning technology and a cloud-first mentality that drives innovation, overcomes obstacles and gives you the confidence that we are not simply chasing your IT budget.

That's what we mean by **connecting the dots**.

The future of IT is in the cloud so it's up to us to deliver smart, secure, scalable and sustainable cloud-centric solutions that are ready when you are.

KONC!SE SOLUT!ONS
CONNECTING THE DOTS
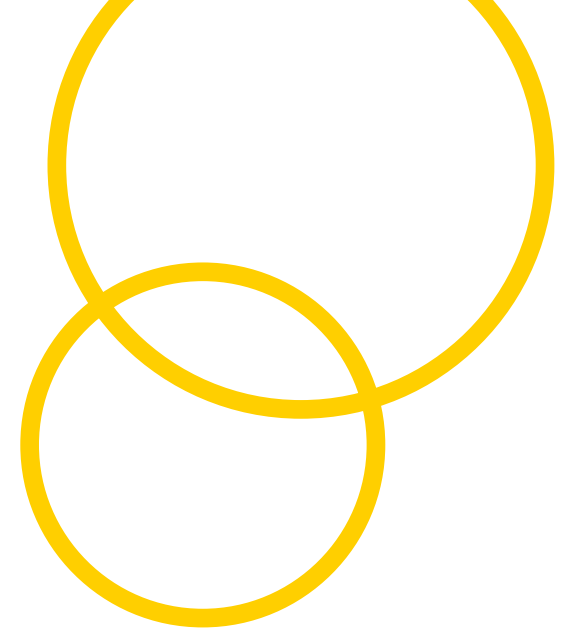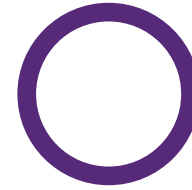
ABOUT CENSORNET

Think of Censornet as another member of your team — one that's faster, smarter, and safer than humanly possible. Our autonomous, integrated cloud security gives mid-market organisations the confidence and control of enterprise-grade cyber protection.

Processing over a billion threats per day, we secure your entire IT environment, freeing you up to concentrate on the business. We protect millions of users across the globe and more than 1,500 customers from a range of cyber security threats.

BENEFICIAL INTELLIGENCE

At Censornet, we believe not so much in artificial intelligence, but in beneficial intelligence. Just because we can do something doesn't mean we should.

EXPOSING THE HYPE: THE ROLE OF AI IN CYBER SECURITY

IT leaders are defending their systems in the face of increasingly sophisticated security attacks. Cybersecurity Ventures expects global cybercrime costs to grow by 15% per year over the next five years, reaching $10.5 trillion USD annually by 2025, up from $3 trillion USD in 2015.[1] Hackers are probing systems for weak spots at scale, corrupting and even poisoning data to obscure their activities. At times, it feels relentless. Can AI – as many vendors claim – protect us?

With so much noise in the market about artificial intelligence, it can be hard to tell the facts from the science fiction. In this ebook, we'll explore the potential of AI technologies to stop attacks, eliminate breaches, ensure compliance and address team fatigue.

1. Cybercrime To Cost The World $10.5 Trillion Annually By 2025, Cybercrime Magazine, 13 November 2020
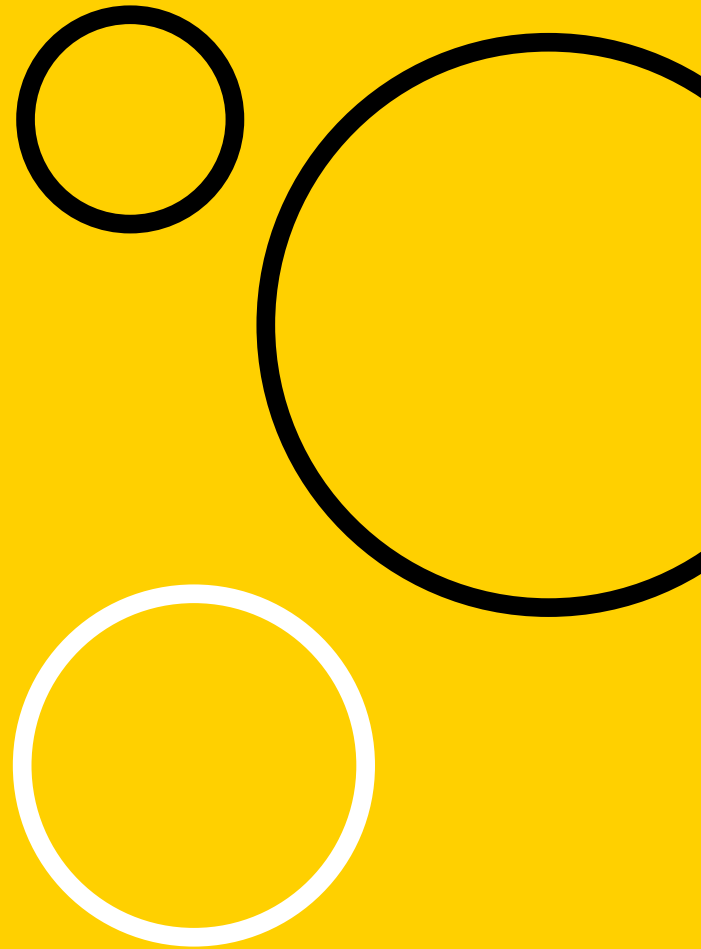
1 Some definitions

When we talk about artificial intelligence as it relates to cyber security, we're really talking about machine learning (ML). This section runs through the different fields within ML.

2 The hype

Many companies claim to have AI capabilities, but the research suggests AI adoption isn't quite what it seems. We evaluate the current state of play.

3 AI use cases in cyber security

Despite the hype, no artificial intelligence can predict what new methods attackers will use next. We'll take a look at how AI can be used within cyber security.

4 Adversarial ML

Attackers understand AI and ML just as well as we do. We'll explore types of attacks including data poisoning, adversarial input and generative adversarial networks.

5 10 steps to effective AI

How can you implement AI that delivers results? Censornet's security experts offer practical advice that applies to any project or service where you're using AI and ML.

6 The Censornet platform

Censornet prevents attacks before they enter the kill chain, by combining a company's core security services – email, web, CASB and MFA – into a single cloud platform.

DEFINITIONS

Before we start talking about AI in earnest, it helps to define what we mean by AI and machine learning. Machine learning is a subfield of AI, alongside fields such as Bayesian statistics and evolutionary algorithms. Even simple rule-based engines could be considered to be a form of AI, albeit perhaps the lowest common form. Then within machine learning are further subfields like supervised, reinforcement and deep learning.

AI

ML

Supervised
Unsupervised
Reinforcement learning

DL

Bayesian statistics	Evolutionary algorithms	Rule-based engines

ML is a subfield of AI, alongside Bayesian statistics, evolutionary algorithms and rule-based engines.

WHAT IS ARTIFICIAL INTELLIGENCE?

" The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages."

Oxford English Dictionary

TYPES OF MACHINE LEARNING (ML)

There are four main types of machine learning, ranging from the simplest – supervised learning – to complex layers of deep neural networks.

1_ SUPERVISED LEARNING

The algorithm 'learns' to classify from a labelled dataset. For example, in image content analysis, algorithms have been trained with millions of images or pictures that are already labelled as safe or not safe for work.

Machine learning has transformed this area of cyber security. By using very large datasets of real-world images, the accuracy of image content analysis has improved from around 50% to well over 90% today.

2_ UNSUPERVISED LEARNING

Here, the algorithm isn't trained in any way. It has no test or development dataset. Instead, it attempts to make sense of an unlabelled dataset by extracting features and patterns, often using k-means clustering.

Unsupervised machine learning is very powerful in user and entity behaviour analysis, and also in intrusion detection systems that are based on ML.

3_ REINFORCEMENT LEARNING

These are self-teaching systems that learn by trial and error using reward signals for feedback, reinforcing actions that deliver good results. It's like how we learn to ride a bike – initially you fall off a lot. Google Deep Mind used reinforcement learning when playing computer games such as the Atari classic 'Breakout'. It made lots of mistakes at first, but quickly improved to beat humans.

4_ DEEP LEARNING

'Deep' simply relates to the number of layers that exist within a DL model. It's typically based on underlying deep neural networks. Data is transformed through multiple layers that progressively extract higher-level features from input data.

A good example again would be image content analysis. Initially, the model will extract general information about the image itself, but subsequent layers will look at things like edges or recognisable shapes, and then human-recognisable shapes such as characters or numbers.

WHERE'S THE POTENTIAL FOR AI IN CYBER SECURITY?

Many advancements have come out of unsupervised and supervised machine learning, but from a cyber security perspective, the field that shows the most potential is deep learning and deep neural networks.

Deep learning has shown the most promise in so many other areas, like digital assistants. But it also requires massive investment, which may be a long time coming given the fragmented nature of the security industry.

HOW WIDESPREAD IS THE HYPE AROUND AI?

NewVantage Partners' 2021 AI survey revealed that adoption of AI initiatives is progressing – 77.8% of companies report AI capabilities to be in widespread or limited production, up from 65.8% the previous year and only 4.1% report no applications of AI in use[2].

So, are we right to talk about hype?

Well, AI isn't new. We've used AI and neural networks to defend against spam email for decades. When it comes to anti-malware, the first AI research papers were published around 1996.

But if we look at how much it's talked about – for example, in vendor earnings calls, there was little mention of AI or ML until around 2013. Everyone was focused on big data.

Then in 2015 and early 2016, there was a huge rise in mentions. Big data was no longer the hot topic. It was now extremely fashionable to talk about AI and machine learning instead.

2. Big Data and AI Executive Survey 2021, NewVantage Partners, Janruary 2021

AI MENTIONS

% of all sentences in the filings of companies in the tech sector which refer to artificial intelligence, by quarter. Source: GlobalData

Today, artificial intelligence mentions have increased 172% since 2016. The percentage of companies which have mentioned artificial intelligence at least once in filings during the past twelve months was 89% compared to 73% in 2016, according to the latest analysis from GlobalData.[3]

For anyone concerned that AI has been overhyped, the world doesn't seem to be moving on. In just the last six years, the number of AI-related publications on arXiv grew by more than sixfold, from 5,478 in 2015 to 34,736 in 2020 .[4]

3. Filings buzz: Tracking artificial intelligence mentions in the tech sector, Verdict, 22 October 2021
4. Artifical Intelligence Index Report 2021, Stanford Unviersity's Human-Centered Artifical Intelligence, March 2021

AI AND CYBER SECURITY

So how does this backdrop of hype around AI translate specifically into cyber security?

The Capgemini Research Institute surveyed 850 senior IT executives across 10 countries and 7 business sectors and conducted in-depth interviews with industry experts.[5] It seems that some of the hype is rubbing off:

- Nearly two thirds think AI will help identify critical threats
- 69% believe AI will be necessary to respond to cyberattacks
- Two thirds plan to deploy AI by 2020
- Organisations are counting on AI to help overwhelmed security analysts improve accuracy, respond faster... and reduce detection and response costs

That's a lot to ask.

5. Reinventing Cyber security with Artificial Intelligence, Capgemini Research Institute, 2019

AI USE CASES IN CYBER SECURITY

According to the Capgemini report, organisations are counting on AI to help security analysts focus on the alerts and events that matter, while reducing detection and response costs. So let's take a look at AI's potential for specific cyber security use cases.

The chart below plots the key use cases for AI, showing the complexity involved in developing each technology against the benefit it can deliver.

CYBER SECURITY USE CASES
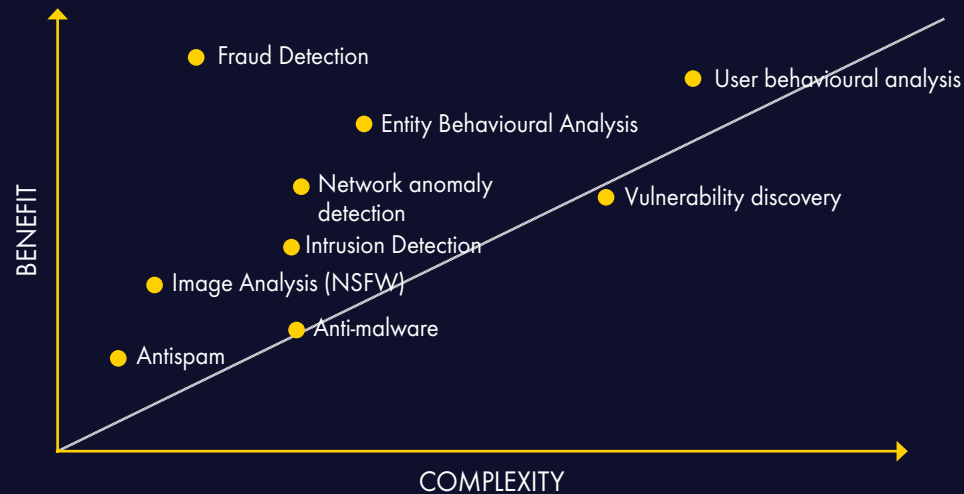
Key use cases for AI plotted by complexity and benefit

ANTISPAM AND ANTI-MALWARE

ML has been used within antispam solutions for decades. Email security first started using ML in the late 1990s. Both behaviour-based antivirus and sandboxing technologies also use ML extensively, but again it's nothing new. One of the earliest papers on using neural networks for virus recognition was published in 1996.

Recently, behaviour-based antivirus has grown in capability and popularity. New packaging techniques can be identified by platforms such as Censornet without a signature update to a traditional signature-based AV solution. With email security, we now have more effective, multiple levels of sandboxing, too.

IMAGE CONTENT ANALYSIS (ICA)

Supervised ML has also transformed ICA (and visual content analysis too – it's the same approach, really, because a video stream is effectively just a series of still frames). Accuracy of ICA has improved from 50% in the late 1990s to more than 90% today, using large-scale training datasets of millions of images against labelled datasets.

FRAUD DETECTION

Fraud detection provides significant obvious benefits – that's why it's towards the top of the chart on the benefit axis. It uses supervised and unsupervised models to identify anomalies in very specific transaction data. Banks use it to identify unusual spending patterns and protect our money.
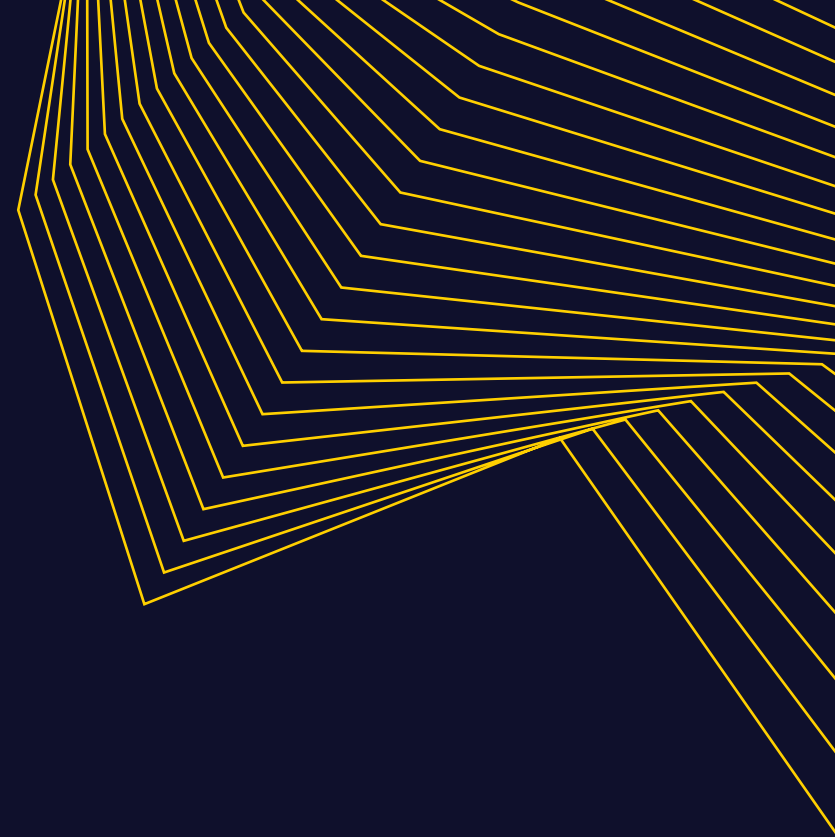
VULNERABILITY DISCOVERY

Vulnerability discovery identifies weaknesses in code before it's released. It works well against both source code and binaries. Over the next few months and years, we should start to see tools emerge that can help with the secure software development lifecycle.

USER AND ENTITY BEHAVIOURAL ANALYSIS

This is one of the more complex areas for AI or ML in cyber security. There's particular potential in user authentication and data access behaviour. The work being done here mostly uses unsupervised learning.

An entity could be a laptop, smartphone or tablet. It could also be a mailbox, where ML would look at the behaviour associated with that mailbox. Or AI could watch the traffic that is requested from or uploaded to a website from an organisation in a set time period. It could also monitor a cloud object, such as a folder within a cloud storage application, and see how often users interact with the folder and how much data is transferred to and from it.

User behaviour and entity behaviour are often combined as user and entity behavioural analysis (UEBA). In cyber security, however, we look at them separately, as user behavioural analysis is far more complex than entity behavioural analysis. That's because it's harder to define what "normal" behaviour is for any one user. It's going to take ML a little while to deliver value with UBA, but ML should deliver value in EBA much sooner.

WHAT CAN'T AI DO?

In cyber security, we talk about Black Swan events. These are completely unforeseen and unpredictable.

Beware of claims of AI or ML being able to make predictions. All models, to some extent, make forecasts. But AI depends on learning from previous datasets. It can't see into the future and anticipate something completely new.

HOW TO DEVELOP AND DEPLOY *ML MODELS*

So how do you effectively develop and deploy a machine learning model into a production environment? Let's take a look at an example that uses supervised machine learning.

1_ First, the model is built, its algorithms are selected, and then it's trained with a training dataset.

2_ Next, a second dataset, usually referred to as the development dataset, is used to confirm that the model has interpreted the training data to a satisfactory level. The engine is then tuned and optimised, again using the development dataset.

3_ A third dataset, called the test dataset, is then used to finalise the model.

4_ Finally, the model is released into a live production environment. It's here that the model is exposed to real life and for the first time gets to deal with live production data.

DEVELOPMENT/ STAGING

PRODUCTION

| Training Data Set |
| Dev Data Set |
| Test Data Set |

⚙️ Train model

🕐 Tune / Optimize

👍 Tune / Optimize
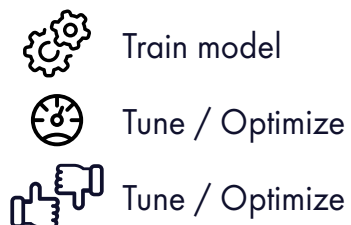
Production Data

The key stages in developing a ML model

Three things to note when developing and deploying machine learning:

1

Development and test datasets must be representative of production data. That means the data needs to be realistic and recent.

2

Datasets should not be available in their entirety in public repositories. There's a risk of bias in the data.

3

Datasets should never be normalised. Anomalies and outliers are necessary. Attackers are unpredictable and often do the opposite of what's expected. You'll need anomalies and outliers in your dataset so your model can learn what an expected anomaly could look like and differentiate that from attackers.

CAN AI HAVE A NEGATIVE EFFECT ON CYBER SECURITY?

If you build the model correctly and you follow the 10 steps to effective AI later in this ebook, then any ML model should be better than its rules-based predecessor. If something does go wrong, however, probably the worst thing that you'll end up with is a false sense of security. That's where careful monitoring and clear process can help you, so you know your AI is doing what you need and expect it to do – and your teams know what to do if it doesn't.
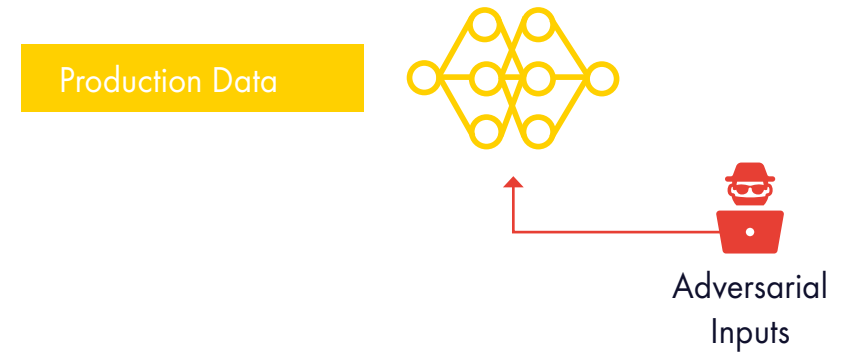
ADVERSARIAL MACHINE LEARNING

As cyber security professionals, we understand the use cases for ML and where it can provide tangible benefits. But we also know that attackers know they can use ML to their advantage, too.

There are two key ways that attackers can target development and runtime models. The first is data poisoning; the second is adversarial input.

ATTACKING ML

| Training Data Set |
| Dev Data Set |
| Test Data Set |

Data Poisoning

⚙ Train model

⏱ Tune / Optimize

👍👎 Tune / Optimize

PRODUCTION

Production Data

Adversarial Inputs

Attackers can target ML with data poisoning during development or adversarial inputs once in production.

DATA POISONING

Data poisoning works well against advanced or complex algorithms. Even with low numbers of poisoned samples, attackers can achieve a significant reduction in test accuracy.

That's why it's best to be cautious about public repositories. Avoid using data that's exclusively in the public domain, because the data is exposed and attackers will actively be trying to poison it.

ADVERSARIAL INPUT

The other form of attack is adversarial input. If you don't manage to poison the data at the development stage, you can still have an impact within the production environment or at runtime. Attackers do this by creating data – examples, actions or activities – that are designed to trick the model in the production environment after it's been deployed. They skew the algorithm's view of what is normal.

Data poisoning against security software that uses AI and ML is thought to be the next big cybersecurity risk. Johannes Ullrich, dean of research of SANS Technology Institute, explained that;

> "One of the most basic threats when it comes to machine learning is one of the attackers actually being able to influence the samples that we are using to train our models."[6]

Adversarial input is a serious concern. Real life examples include researchers discovering methods to alter the appearance of a stop sign so that an autonomous vehicle classified it as a speed limit sign.[7]

To prove the ease of utilising adversarial input, researchers at McAfee tampered with a speed limit sign by simply adding a two-inch strip of black tape. This tricked the Mobileye EyeQ3 camera on a 2016 Tesla Model X and Model S into feeding bad information to the vehicles' autonomous driving features, causing both cars to accelerate to 50mph over the speed limit.[8]

6. The Five Most Dangerous New Attack Techniques, SANS Institute, 10 June 2021
7. Breaking neural networks with adversarial attacks, Towards Data Science, 9 February 2019
7. Model Hacking ADAS to Pave Safer Roads for Autonomous Vehicles, McAfee, 19 February 2020

ATTACKS AGAINST DEEP NEURAL NETWORKS

There's a lot of research into creating adversarial examples that fool Deep Neural Networks (DNNs). Small imperceptible amounts of 'perturbation' or 'noise' can throw these networks way off course.

Here's a famous example. To the human eye, both pandas look identical. The neural network analysing the images identified the original image as a panda with just under 60% confidence. But once a small amount of noise was added, that same neural network was over 99% confident that the image of the panda actually portrayed a gibbon. [8]
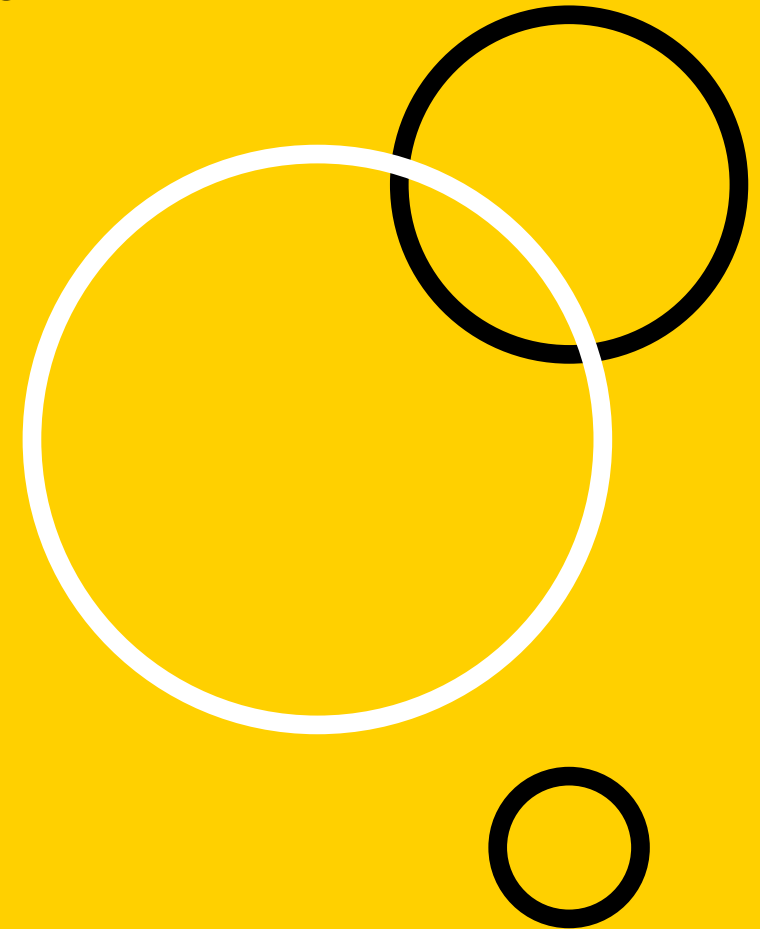

Panda

57.7% confidence

Gibbon

99.3% confidence

A tiny amount of noise tricked the neural network into thinking the panda on the right was a gibbon.

8. Explaining and Harvesting Adversarial Examples, arXiv:1412.6572

ML AND FACE RECOGNITION

In another example, researchers at Carnegie Mellon University used printed eyeglasses to evade being recognised or impersonate another individual.

On the left, the man in bright printed glasses has fooled the facial recognition system into believing he's Milla Jovovich. In the middle, superimposed glasses have convinced the system that Reese Witherspoon is actually Russell Crowe. On the right, the man in glasses has tricked the system into thinking he's Carson Daly.

Impersonating Milla Jovovich; Reese Witherspoon and Russell Crowe; impersonating Carson Daly

GENERATIVE ADVERSARIAL NETWORKS (GANS)

It's difficult and complex to generate adversarial inputs, and each is specific to the particular model that you're trying to attack at runtime. So networks called Generative Adversarial Networks, or GANs, have emerged to help.

People use GANs every day. They can be used to:
• Power filters in photography apps
• Auto-generate and colour anime characters
• Transform photographs into images in the style of Monet or Van Gogh
• Create high-resolution images from low-res originals or synthetic maps
• Age our faces
• Blend images to generate video or music

But GANs can also be used for malicious purposes.

AI can generate targeted tweets six times faster and also have twice the success than human-generated tweets. We're seeing more and more deepfake videos influence public opinion and the cost is falling all the time. In August 2019, the Wall Street Journal broke the first case of deepfake voice fraud. AI had been used to deepfake a German CEO's voice and persuade the receiver of the call to transfer $243,000 to a Hungarian supplier.[9]

Right now, we're in a perfect storm where computer processing power is at an accessible price point to analyse the explosion in the amount of data that we generate. AI models and tools are freely available. There are courses on Coursera and Udacity, frameworks like TensorFlow, even dedicated development platforms like Jupyter Notebook.

And if we have access to all these tools and educational resources, so do the attackers.

But the security industry is – some would say fortunately – highly fragmented. Few vendors exist with sufficient resources to pour tens or hundreds of billions of dollars into AI for cyber security.

Cyber security use cases will therefore take longer to develop than the digital assistants most of us already have in our homes.

9. Fraudsters used AI to mimic CEO's voice in unusual cybercrime case, Wall Street Journal, 30 August 2019

10 STEPS TO EFFECTIVE AI

These steps apply however you're using AI in cyber security.

1
Take at least a partial 'black box' approach

If an attacker has access to the full ML model or application, reverse engineering is easier. Black box cloud deployments limit the ability to abuse the model without being detected.

2
Use a lot of representative test data

Make sure that test data is realistic and recent – and the more you have of it, the better. Large datasets will refine your model and raise its accuracy.

3
Expect the unexpected

Don't normalize or filter out the anomalies and the outliers in your test data, because when you move into production at runtime, attacks are nearly always going to be anomalies.

4
Don't rely on public domain test data alone

By all means, use test data that's in public repositories, but protect yourself in case it's poisoned: supplement it wherever possible with private data.

5
Monitor for suspicious new training data

There's so much evidence of data poisoning taking place in the real world. Monitor for any suspicious new training data that gets added to the test or dev datasets.

6

Check newly trained model quality against previous models

When you've trained a new model, make sure that you compare its quality with previous models in case there are any unexpected changes.

7

Measure bias, variance and error rates throughout the development lifecycle

It's never good if you can introduce bias into the model so that it favours one outcome over another. You also want minimal rates of variance and errors.

8

Do not depend on ML as a single layer of defence

ML will be compromised: it's a case of when, not if. Never depend on ML as a single layer of defence. It's been true in security for many years: you need multiple layers.

9

Anticipate adversarial inputs and use adversarial training

Put yourself in the role of an attacker. Think about the adversarial inputs that may be used and train the model against them so it's more robust when it encounters them at runtime.

10

Stay ahead of new developments

AI is a field that is moving so fast. You've got to stay on top of new developments, techniques, algorithms and so on, because you can bet that the attackers are.

HOW CENSORNET & KONCISE CAN HELP

We believe in Beneficial Intelligence – using AI in a thoughtful, positive way. Our unique cloud security solution uses advanced machine learning to give you automated attack prevention.

Censornet combines your organisation's core security services – email security, web security, CASB and MFA – on a single, tightly integrated cloud platform. This gives you full visibility and advanced protection from the fast-moving multi-channel threats that are so common today.

AUTONOMOUS INTEGRATED CLOUD SECURITY

E. EMAIL   **W.** WEB   **C.** CASB   **M.** MFA

Enable traditionally silo'd products to share and react to security events and state data whilst leveraging world class threat intelligence. Prevent attacks before they enter the kill chain.

Censornet's integrated cloud platform uses ML to prevent attacks before they enter the kill chain.

HOW IT WORKS

Censornet reacts at machine speed to stop attacks before they even enter the kill chain.

Here's a real-world example. Imagine an employee receives a phishing email into their Office 365 account. The email security module in our platform extracts the malicious URL from the email message body and shares it with our Autonomous Security Engine (ASE). Web security – also integrated with ASE – automatically adds it to the block list for all users in the organisation.

If a user uploads a file containing personal, sensitive or regulated data to a cloud app like Dropbox, then the CASB service will share the hash of the file with ASE. ASE then shares that hash out, so if the user then tries to exfiltrate the same file via webmail, Slack or email, ASE automatically blocks those attempts.

CONSTANTLY UPDATED

We continually update ASE with multiple commercial, open source and government-provided threat intelligence feeds.

Core services in turn automatically update to protect against new domains and IPs that are 'phishy' or 'spammy', part of 2C/CnC or botnet infrastructure, distributing malware, serving up fake login pages and so on.

Censornet's adaptive security then learns and evolves in real time as the threat landscape changes.
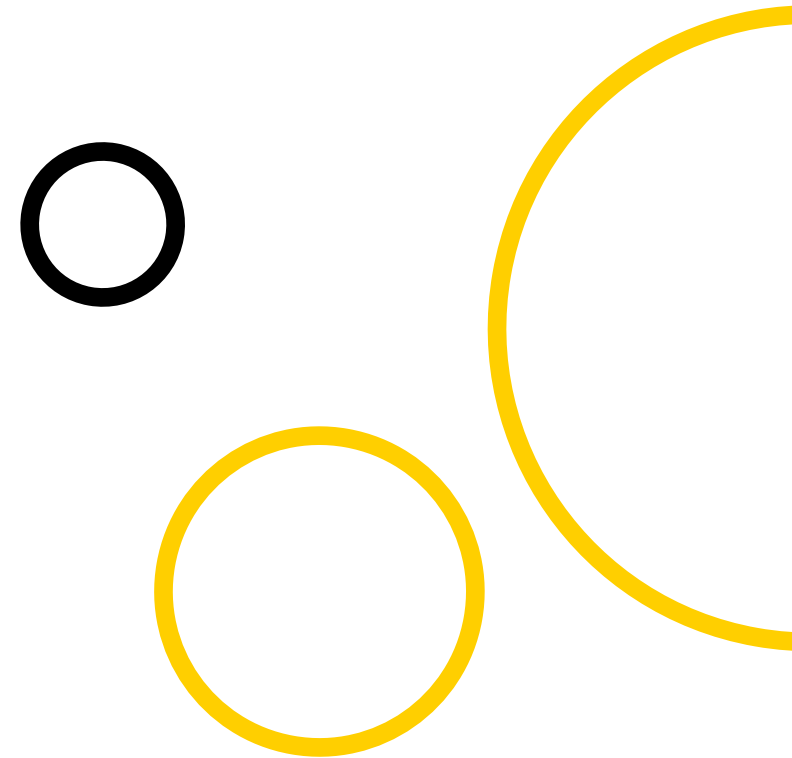
WHAT'S THE ALTERNATIVE?

Compare this seamless and automated workflow and response with one where you have siloed email, web, cloud and authentication products.

If that's the case, each step of a multi-channel attack would be executed against defences that are isolated and uncoordinated. Each one would need a manual response to an alert. That means slower response times – and a dramatically higher probability that an attack will succeed.

To find out more about how Censornet can keep your organisation safe 24/7, give us a call on **+44 (0) 845 230 9590** or drop us a line at **censornet.com/contact**

We'd love to hear from you.

censornet.