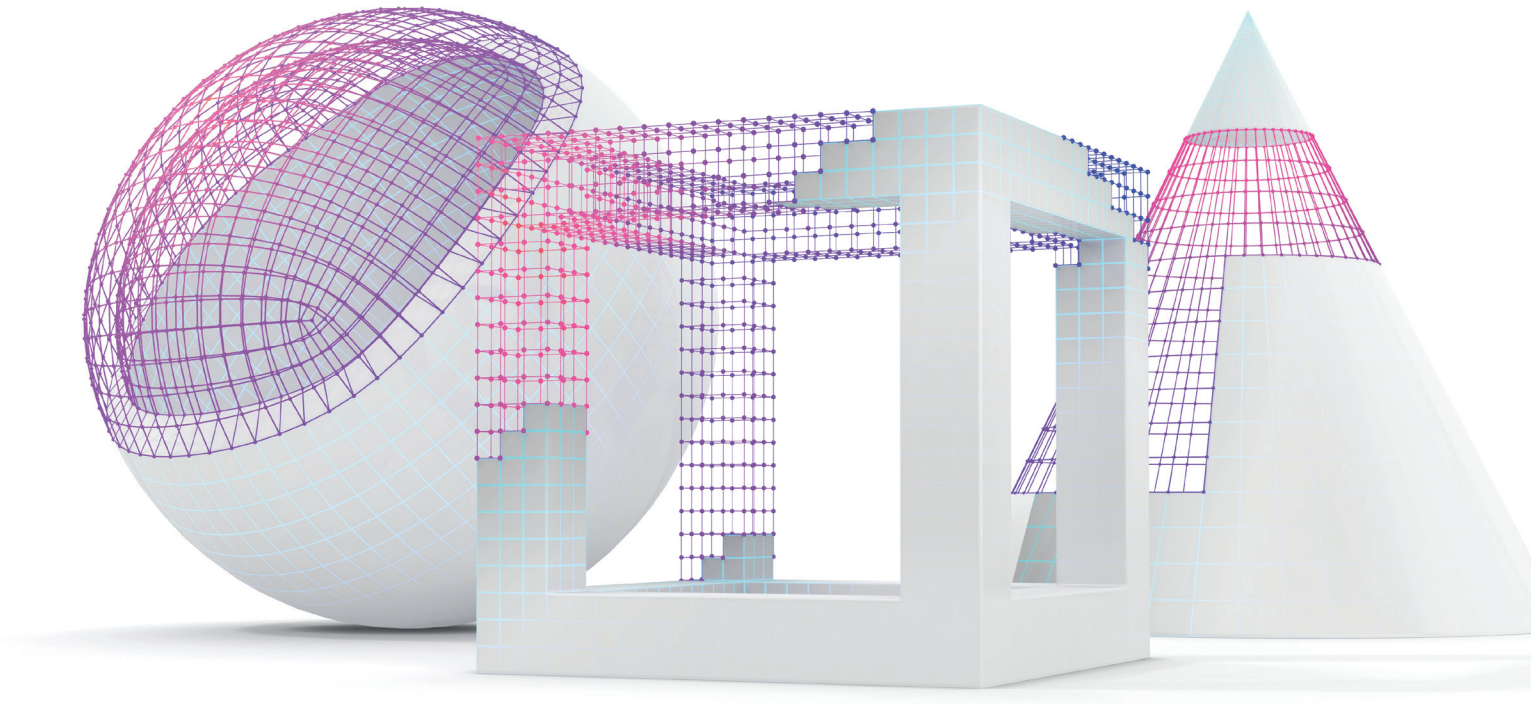
Accenture Labs

Know your threat

AI IS THE NEW ATTACK SURFACE



ARTIFICIAL INTELLIGENCE (AI) IS INCREASINGLY PERVASIVE.

This creates huge advantages for individuals, organizations and society. It also presents major risks. As "adversarial AI" has emerged over the past five years, we've seen more and more attackers exploiting machine learning models and using them to benefit their own interests.

As they've evolved to capture ever higher levels of complexity, machine learning models have lost their interpretability. This opens new avenues—across continually expanding attack surfaces—for motivated attackers to exploit.

In this world, the creation of robust, secure AI is a top priority for all organizations. Otherwise AI models can and will be exploited, sometimes with potentially disastrous consequences.

Once, organizations only needed to secure the infrastructures underpinning their AI models. Now they must secure the models themselves.

> The bottom line? Algorithms are the new front-line for cybersecurity teams.

Artificial intelligence, machine learning… and the rise of adversarial AI

The definition of artificial intelligence has changed over the years. Today, it's most often used as an overarching term to describe many different types of problem-solving methods through analytics and automation.

AI includes diverse technologies—from robotics to rule automation to machine learning. Each of them has been used increasingly by organizations in the past decade. But one technology, above all, has been in the forefront of value creation: machine learning.

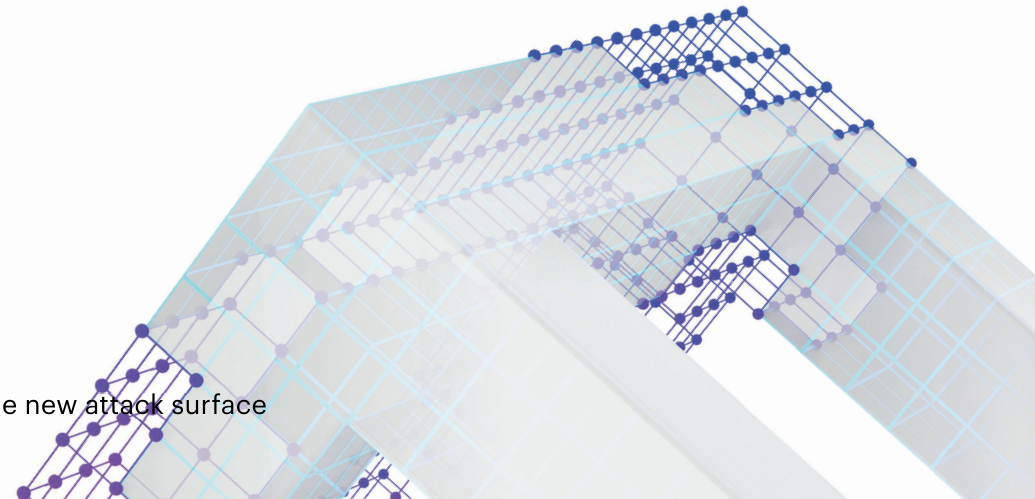
A broad field originating from statistics and operations research, machine learning is, at its core, a method for predicting outcomes from sets of data. It does this by creating models that automatically infer patterns in the data and use those patterns to make decisions.

The recent rise of cheap computational power and abundance of collected data has democratized this technology by allowing practitioners to develop increasingly complex models of behavior at low cost.

Most of these machine learning models are black boxes. That's to say, as the accuracy and complexity of these models has continued to grow, many of the behaviors they capture defy any comprehensive human understanding. It's this complexity and unexplainable model behavior that create the potential for exploitation.

It's why adversarial AI has become such a potent threat: if an adversary can determine a particular behavior in a model that is unknown to its developers, they can exploit that behavior for potential gain.

Adversarial AI attacks succeed, in most cases, by predicting the decisions machine learning models will make and then manipulating subsequent sets of data to produce the attacker's desired outcomes—rather than the correct decisions. "Poisoning attacks" can also take place, where the machine learning model itself is manipulated.

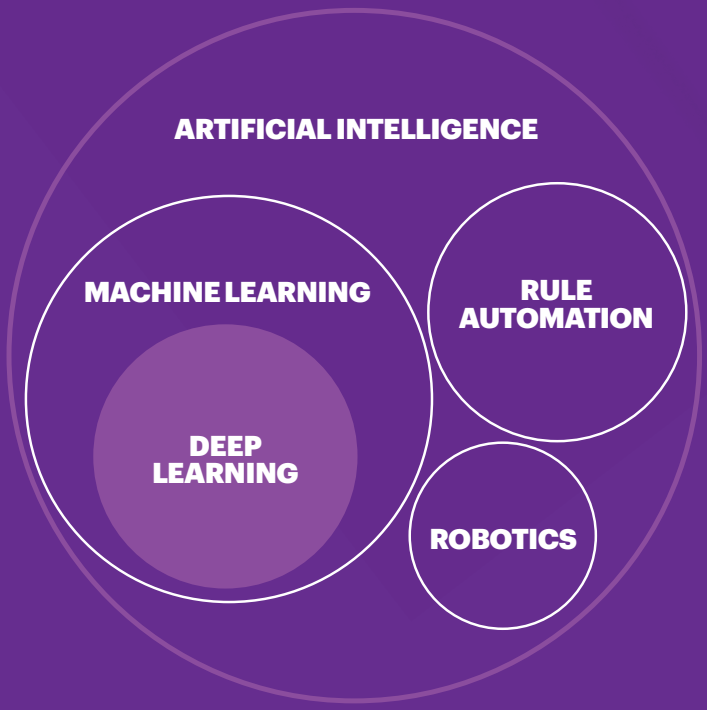


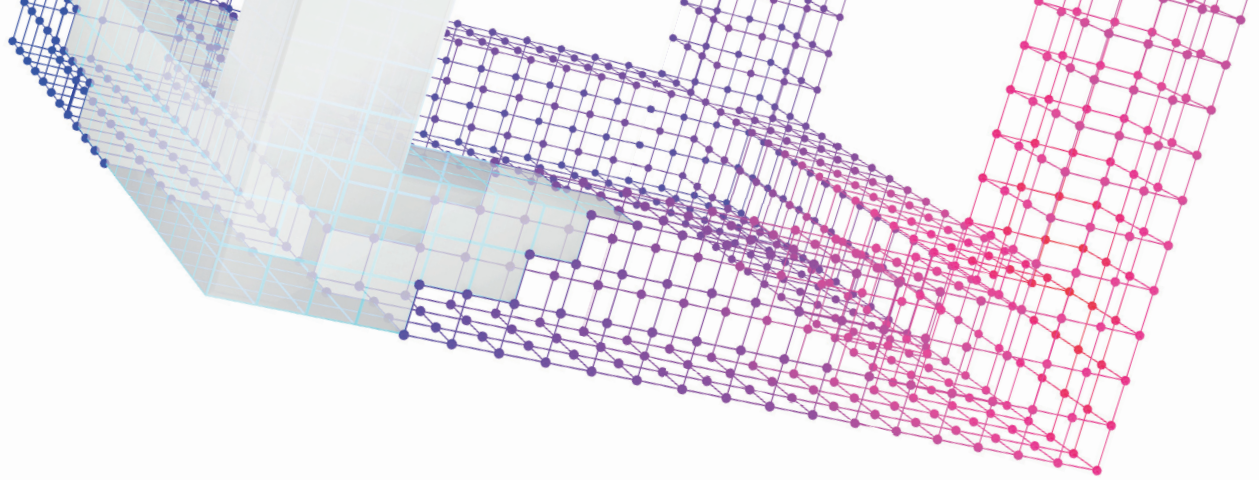
WHAT IS DEEP LEARNING?

> Deep learning is a type of machine learning. But unlike regression-based methods, it uses complex networks of layers of “neurons” or “units”.

These make very simple calculations of their inputs, before passing the results of those calculations onto the next layer. The final layer of a neural network will be the model’s result. Based on how far this is from the actual result of the training instance, the algorithm will go back and adjust every neuron to make the calculated result closer. Over time, with large training sets, the algorithm will “learn” what the desired results of a particular problem are and be able to replicate that behavior on unlabelled datasets.

Deep learning is extremely useful for solving problems that are difficult to replicate with code. But it often produces models that resemble “black boxes” and lack interpretability, which make them vulnerable to exploitation.





Adversarial AI... know your enemy

Adversarial AI causes machine learning models to misinterpret inputs into the system and behave in a way that's favorable to the attacker. To produce the unexpected behavior, attackers create "adversarial examples," often resembling normal inputs, but instead meticulously optimized to break the model's performance.

Attackers typically create these adversarial examples by developing models that repeatedly make minute changes to the model inputs. Eventually these changes begin to stack up, causing the model to become unstable and make inaccurate predictions on what look like normal inputs.

With enough of these small modifications, attackers can dictate a model's response for any input, using this to provide the outcomes they want. These outcomes can be extremely diverse, from defeating content filters right through to causing undetectable catastrophic failures.

While state-of-the-art modeling techniques like random forests and support vector machines have also been shown to be vulnerable, the high complexity of deep learning models makes them particularly susceptible to adversarial attack.

Adversarial AI and deep learning have a symbiotic relationship with each other. Deep learning's effectiveness is built on the extremely large number of interactions between individual neurons in a network.

It's these interactions that create the extremely high complexity that enables deep learning models to predict behavior so accurately. But this complexity also comes with a downside, providing an entry point for adversarial attacks.

Creating adversarial examples is a complex undertaking. The best way is to use deep learning to learn how to manipulate inputs into the attacked system. A branch of deep learning models, "Generative Adversarial Networks (GANs)", do just that. Most adversarial attacks use GANs to create adversarial examples, optimized to fool an attacked model and produce the desired outcome.

ADVERSARIAL AI: A HISTORY

> Attackers have been generating adversarial examples for decades. But until recently, these had to be created through a manual process.

Some of the earliest attempts focused on spam filters. To start with, spammers created messages selling various products and sent them to multiple email addresses. The first spam filters honed in on specific keywords in those emails to block and reroute them.

Attackers evolved, becoming increasingly creative in generating adversarial examples. Filters also grew more sophisticated to the point that, today, they're complex artificial intelligence models for spam detection. But the flow of spam hasn't slowed down. There are still continual attempts to generate spam that eludes filters, either through hand-crafted messages or AI.

Another example of adversarial AI is "Search Engine Optimization (SEO)". Outside actors learned that they could manipulate the algorithms used by search engines to rank the popularity of websites in search results.

They used this knowledge to raise the popularity of their sites and drive additional traffic to them. Search engine companies now routinely update their algorithms to prevent SEO boosting from impacting user search results.

A clear and present threat

Researchers in adversarial AI have created proof-of-concept attacks against a number of core technologies like computer vision, OCR (Optical Character Recognition) and malware detection. A brief examination of some of these highlights the methodologies and desired outcomes behind these attacks.

Computer vision

Many recent advances in computer vision have been enabled by deep learning—from classifying image content and creating decision-making processes for self-driving cars to recognizing objects in surveillance feeds.

Image content/classification is one of the most researched areas of adversarial AI. A typical attack in this space generates an adversarial example which is given to a machine learning model. Because of manipulation, the model misinterprets the content of the image and misclassifies it.

In this way, an attacker can tailor the expected behavior of an algorithm to achieve a number of outcomes. And in self-driving car use cases, researchers have created adversarial examples that can cause accidents.

Widely used by organizations to extract text from images, OCR software is another area at risk of attack. Proof-of-concept adversarial attacks have caused OCR systems to misread the information from images that is then translated to text. Fraud use cases represent one of the broadest attack vectors (online banking apps could be exploitable).

Natural language processing (NLP)

Recent research shows that applications of deep learning in NLP are also vulnerable to adversarial attacks. Unlike images, which are usually scaled to have continuous pixel intensities, text data is largely discrete. This makes optimization for finding adversarial examples more challenging.

Adversarial examples in this space focus on inducing misclassifications through changes that maintain semantic similarity (sentences with similar meanings are close to each other), or making changes that maintain syntactic similarity (sentences are structured the same).

The objectives of these adversarial attacks are varied, and could include subversive manipulation of the algorithms that determine sentiment, gather intelligence, or filter for spam and phishing.

Industrial control systems
To reduce computational complexity, many control systems make estimations and approximations. This simplification means that some interactions will not be captured in the control equations. By creating GANs that make minor manipulations (that may go unseen by human operators) to control systems' inputs, attackers can cause unexpected behaviors that create a wide array of outcomes—from simple system degradation, to increased wear-and-tear, to catastrophic failure.

The AI attack surface... a new area of vulnerability

The majority of organizations' current investment into security is dedicated to securing the hardware and software attack surface. These include patching vulnerabilities, static and dynamic analysis of production codes, and OS hardening.

This overlooks a key point: adversarial AI targets areas of the attack surface that have never previously had to be secured, the AI models themselves. From now on, organizations need to include these in their security budgets—or risk them being exploited by attackers.

Securing an AI model requires different skills and toolsets than securing code. In large part, that's because it's impossible to test every combination of inputs for an AI model (the number of values taken by a single variable can be infinite).

Until recently, data scientists addressed this problem by using sensitivity and robustness testing. But these tests are used to test for stability on random inputs, not specific combinations of inputs engineered to trigger unexpected behavior. And they're more likely to fail to predict behavior in more complex models.

To ensure their AI models are robust enough to withstand exploitation, organizations must take advantage of adversarial AI counter-measures and emerging practices. The AI attack surface is an entirely new area of infrastructure that has to be secured. Security practices must adapt to accommodate it.

CREATING ROBUST, SECURE AI

So how can the AI attack surface be comprehensively protected?

It's a complex challenge and organizations will need to combine multiple approaches to ensure robust, secure AI. Here's a brief introduction (we'll look at them in greater depth in a follow-up paper):

RATE LIMITATION

by rate-limiting how quickly individuals/systems can submit a set of inputs to a system, organizations can increase the effort it takes to train their models. That's a major deterrent to adversarial attackers.

INPUT VALIDATION

focusing on what's being put into AI models. By making small modifications to an adversarial example, it's often possible to "break" its ability to fool a model.

ROBUST MODEL STRUCTURING

the structuring of machine learning models can provide some natural resistance to adversarial examples.

ADVERSARIAL TRAINING

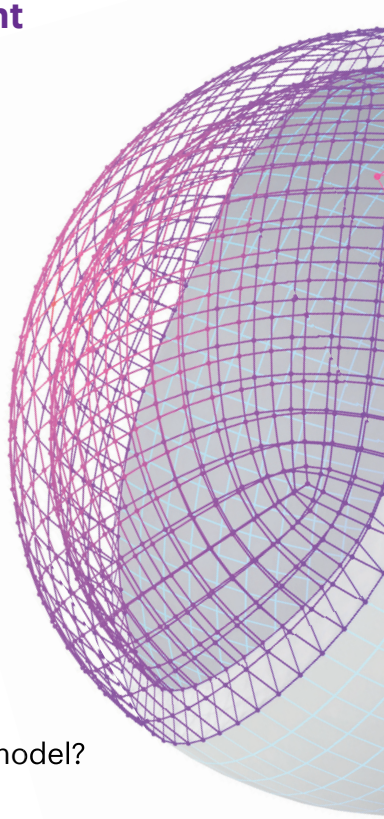
if enough adversarial examples are inserted into data during the training phase, a machine learning algorithm will eventually learn how to interpret them.

National Security and AI

Governments around the world are announcing national strategies for AI and are calling it an economic and national security imperative. For example, in February, the White House issued an Executive Order, outlining the U.S.'s national strategy for artificial intelligence (AI Initiative). The Initiative includes, among other things, necessary research and development investments to foster public trust and confidence in AI technologies. As part of the Initiative the Defense Advanced Research Projects Agency (DARPA) is working on a project to develop technologies to defend against adversarial AI.

Securing AI models... getting started

Even though AI attack surfaces are only just emerging, organizations' future security strategies should take account of adversarial AI, with the emphasis on engineering resilient modelling structures and strengthening critical models against attempts to introduce adversarial examples.

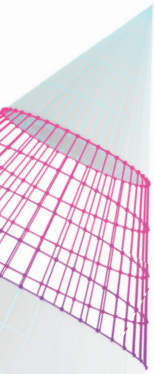


Immediate priorities are to:

01 **Conduct an inventory to determine which business processes leverage AI, and where systems operate as black boxes**

02 **Gather information on the exposure and criticality of each AI model discovered in Step 1 by asking:**

- Does it support business-critical operations?
- How opaque/complex is the decision-making for this process?
- Is the process exposed to the outside world?
- Can customers create their own inputs and get results from the model?
- Are there similar open-source models to this process?
- What potential outcomes could an attacker drive from this model?



03 **Prioritize plans for highly critical and highly exposed models:**

- Using the information gathered in Step 2, prioritize each model and create a plan for strengthening models that support critical processes and are at high risk of attack
- To support prioritization, create trade-off matrices that weigh criticality vs the risk and exposure of each model.

CONTACTS

Malek Ben Salem, Ph.D.
Technology Research Sr Principal
Cyber Security R&D Lead, Accenture Labs
Malek.ben.salem@accenture.com

Jean-Luc Chatelain
Managing Director,
CTO Applied Intelligence
jean-luc.chatelain@accenture.com

Josh Ray
Managing Director,
Global Cyber Defense Lead,
iDefense GM, Accenture Security
joshua.r.ray@accenture.com

CONTRIBUTORS

Louis DiValentin
Technology Research Principal
louis.divalentin@accenture.com

Iman Zabett
Technology R&D Senior Analyst
iman.zabett@accenture.com

Hemanth Ravikumar
Researcher
hemanth.ravikumar@accenture.com

ABOUT ACCENTURE

Accenture is a leading global professional services company, providing a broad range of services and solutions in strategy, consulting, digital, technology and operations. Combining unmatched experience and specialized skills across more than 40 industries and all business functions—underpinned by the world's largest delivery network—Accenture works at the intersection of business and technology to help clients improve their performance and create sustainable value for their stakeholders. With 469,000 people serving clients in more than 120 countries, Accenture drives innovation to improve the way the world works and lives. Visit us at **www.accenture.com**.

ABOUT ACCENTURE LABS

Accenture Labs incubates and prototypes new concepts through applied R&D projects that are expected to have a significant strategic impact on Accenture and its clients. Our dedicated team of technologists and researchers work with leaders across the company and business partners to invest in, incubate and deliver breakthrough ideas and solutions that help our clients create new sources of business advantage.

Accenture Labs is located in seven key research hubs around the world: San Francisco, CA; Washington, D.C.; Dublin, Ireland; Sophia Antipolis, France; Herzliya, Israel; Bangalore, India; and Shenzhen, China; and 25 Nano Labs. The Labs collaborates extensively with Accenture's network of nearly 400 innovation centers, studios and centers of excellence located in 92 cities and 35 countries globally to deliver cutting-edge research, insights and solutions to clients where they operate and live. For more information, please visit **www.accenture.com/labs.**